Analytics for Accelerating Biomedical Innovation

by

Kien Wei Siah

B.Eng., National University of Singapore (2015)
S.M., Massachusetts Institute of Technology (2017)

Submitted to the
Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Electrical Engineering and Computer Science

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

February 2021

Author .....
Department of Electrical Engineering and Computer Science
December 17, 2020

Certified by .....
Andrew W. Lo
Charles E. and Susan T. Harris Professor, Sloan School of Management
Thesis Supervisor

Accepted by .....
Leslie A. Kolodziejski
Professor of Electrical Engineering and Computer Science
Chair, Department Committee on Graduate Students

Analytics for Accelerating Biomedical Innovation

by

Kien Wei Siah

Submitted to the Department of Electrical Engineering and Computer Science
on December 17, 2020, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy in Electrical Engineering and Computer Science

Abstract

Despite the many breakthroughs in biomedical research and the increasing demand for new drugs to treat unmet medical needs, the productivity of research and development in the pharmaceutical industry has been steadily declining for the past two decades and is at its lowest level today. Traditional sources of financing in biopharma are no longer compatible nor aligned with the new realities of biomedical innovation, a process which has become more challenging, complex, expensive, time-consuming, and risky in the past twenty years. This has led to an outflow of capital from the biopharma industry, creating an ever-widening gap in funding between early-stage basic biomedical research and late-stage clinical development, where many promising academic discoveries fail not because of bad science but due to financial reasons. In this thesis, we explore the use of data analytics to facilitate biomedical innovation with a particular emphasis on the mismatch between the risk characteristics of biomedical projects and the risk preferences of biopharma investors.

We begin with a brief introduction of the challenges faced by the biopharma industry in Part I. In Part II, we focus on analytics in the context of clinical trials. First, we develop analytics for precision medicine in non-small cell lung cancer, an emerging area of innovation in disease treatment with the advent of human genome sequencing. Next, we train and validate predictive models for estimating the probability of success of drug development programs. By providing greater risk transparency, our models can help facilitate more accurate matching of investor risk preferences with the risks of biomedical investment opportunities, thus increasing the efficiency of capital allocation. Finally, we turn our attention to the ongoing COVID-19 (coronavirus disease 2019) pandemic. We propose a systematic framework for quantitatively assessing the potential costs and benefits of different vaccine efficacy trial designs for COVID-19 vaccine development, including traditional and adaptive randomized clinical trials, and human challenge trials (HCTs). Our results contribute to the current ethical debate about HCTs by identifying situations where HCTs can provide greater social value versus non-challenge development pathways, and are thus justifiable.

In Part III, we explore new business models to address the dearth of funding for translational medicine in the valley of death. In view of the increasingly critical role

that academic institutions play in the biotechnology industry, we develop a systematic framework for tracking the financial and research impact of university technology licensing in the life sciences using the Massachusetts Institute of Technology as a case study. Next, we investigate the use of a recently proposed megafund structure for financing early-stage biomedical research. We extend the existing model to account for technical correlation between assets in the underlying portfolio, thus allowing us to evaluate the tail risks of the megafund more accurately. We show that financial engineering techniques can be used to structure the megafund into derivatives with risk-reward characteristics that are attractive to a broad range of investors. This allows the fund to tap into a substantially larger pool of capital than the traditional sources of biopharma funding. In the last part of the thesis, we further extend the megafund framework to include adaptive clinical trial designs, and demonstrate the economic viability of using the megafund vehicle to finance and accelerate drug development for glioblastoma, a disease with very few treatment options, low historical probabilities of success, and huge unmet need.

Thesis Supervisor: Andrew W. Lo
Title: Charles E. and Susan T. Harris Professor, Sloan School of Management

Acknowledgments

I am grateful to my thesis supervisor, Professor Andrew W. Lo, for his guidance and support over the past five years. I am deeply inspired by his wisdom, vision, and passion for research with real-world impact. He has been and always will be a role model in my career.

I would also like to thank my thesis committee, Professor Martha L. Gray and Dr. Sean Khozin, and RQE committee, Professor Peter Szolovits and Professor John V. Guttag, who have provided invaluable feedback on my work and encouraged me to see the big picture.

I am thankful to everyone in the MIT Laboratory for Financial Engineering: Jayna Cummings, Crystal Myler, Mavanee Nealon, and Kate Lyons for everything they have done to support our research; Chi Heem Wong, Samuel Huang, Qingyang Xu, Shomesh Chaudhuri, Zied Ben Chaouch, and Manish Singh for the many useful discussions and joint work on several projects. They have made my PhD journey an amazing and enjoyable experience.

I was fortunate to have the opportunity to work with many collaborators and co-authors during my time at MIT: David Aron, Donald Berry, Scott Berry, Christine Blazynski, Meredith Buxton, John Frishkopf, Olga Futer, Mark Gordon, Jerry Gupta, Peter Hale, Michael Hay, Leah Isakov, Nicholas Kelley, Sean Khozin, Grace Lindsay, Jeff Lura, Lita Nelsen, Lesley Millar-Nicholson, Kirk Tanner, and Richard Thakor. I have learned so much about drug development and the biopharma industry from each and everyone. I'm also grateful to the MIT Laboratory for Financial Engineering and the Rockefeller Foundation for funding support. The views and opinions expressed in this thesis are solely my own, and do not necessarily represent the views and opinions of any institution or agency, or any of the individuals acknowledged above.

Finally, being far from home has been a great challenge. I would like to thank my friends and family for their unwavering love and support back in Singapore, without which none of this would be possible.

Contents

I Introduction **19**

1 Challenges of Biomedical Innovation **21**

1.1 Introduction 21

1.2 Thesis Contributions 23

1.2.1 Clinical Trial Analytics 23

1.2.2 New Business Models 25

II Clinical Trial Analytics **29**

2 Predictive Models for Patient Outcomes in Lung Cancer **31**

2.1 Introduction 32

2.2 Data 33

2.2.1 Study Population 33

2.2.2 Tumor Response Data 35

2.2.3 Longitudinal Tumor Size Data 43

2.2.4 Survival Data 44

2.3 Methods 48

2.3.1 Stochastic Model for Tumor Growth 48

2.3.2 Machine Learning Models for Objective Response 51

2.3.3 Statistical Models for Survival 52

2.4 Results 53

2.4.1 Tumor Response 53

2.4.2 Survival . . . 57
2.5 Discussion . . . 64

3 Predictive Models for Drug Development Programs 71
3.1 Introduction . . . 72
3.2 Data . . . 75
3.2.1 Summary Statistics . . . 75
3.2.2 Missing Data . . . 80
3.3 Methods . . . 81
3.3.1 Statistical Imputation . . . 83
3.3.2 Machine Learning Models . . . 88
3.4 Results . . . 90
3.4.1 Imputation Versus Listwise Deletion . . . 90
3.4.2 Predicting Drug Approvals . . . 96
3.4.3 Predictions Over Time . . . 99
3.5 Discussion . . . 104

4 Cost/Benefit Analysis of Vaccine Trial Designs for COVID-19 115
4.1 Introduction . . . 116
4.2 Simulation Framework . . . 118
4.3 Vaccine Efficacy Trial Designs . . . 119
4.3.1 Traditional Randomized Clinical Trial . . . 119
4.3.2 Optimized Randomized Clinical Trial . . . 121
4.3.3 Adaptive Randomized Clinical Trial . . . 122
4.3.4 Human Challenge Trial . . . 122
4.4 Efficacy Analysis . . . 124
4.4.1 Fixed-Duration Clinical Trial . . . 124
4.4.2 Superiority-by-Margin Testing . . . 126
4.4.3 Adaptive Clinical Trial . . . 127
4.5 Epidemiological Model . . . 131
4.6 Cost/Benefit Analysis Framework . . . 133

4.7 Results . . . 134
4.8 Discussion . . . 136
4.9 Conclusion . . . 144

III New Business Models 145

5 Impact of University Technology Licensing: A Case Study of MIT 147
5.1 Introduction . . . 148
5.2 Data . . . 152
5.3 Measures of Impact . . . 152
5.3.1 Orange Book Citations . . . 152
5.3.2 Initial Public Offerings . . . 154
5.3.3 Mergers and Acquisitions . . . 157
5.3.4 Research and Development Pipeline . . . 157
5.3.5 Drug Approvals . . . 161
5.3.6 Intellectual Property . . . 163
5.4 Discussion . . . 165
5.5 Conclusion . . . 169

6 Financing Correlated Drug Development Projects 171
6.1 Introduction . . . 172
6.2 Methods . . . 173
6.2.1 Simulation Framework . . . 173
6.2.2 Parameters . . . 177
6.2.3 Gaussian Copula . . . 179
6.2.4 Impact of Correlation on Tail Risk . . . 183
6.3 Results . . . 187
6.3.1 Simulation . . . 187
6.3.2 Sensitivity Analysis . . . 192
6.4 Discussion . . . 202

7 Financing Treatments for Glioblastoma **205**

7.1 Introduction . . . 206

7.2 Parameters . . . 207

7.2.1 Portfolio . . . 208

7.2.2 Probability of Success, Cost of Development, and Duration . . . 212

7.2.3 GBM AGILE . . . 212

7.2.4 Correlation . . . 216

7.2.5 Profitability of an Approved Compound . . . 216

7.3 Results . . . 218

7.3.1 Baseline . . . 218

7.3.2 Sensitivity Analysis . . . 222

7.4 Discussion . . . 225

7.5 Conclusion . . . 226

IV Conclusion **227**

8 Summary of Findings **229**

V Appendices **233**

A Supplement to Chapter 2 **235**

A.1 Scaling RECIST Measurements . . . 235

A.2 Case Studies of Longitudinal Data . . . 237

A.3 Features for Predictive Models . . . 240

B Supplement to Chapter 3 **243**

B.1 Data Preprocessing . . . 243

B.2 Multiple Imputation . . . 248

B.2.1 Imputation . . . 248

B.2.2 Analysis . . . 249

B.2.3 Pooling . . . 249

B.3 Imputation Versus Listwise Deletion . . . 251
B.3.1 Simulating Missingness . . . 259
B.4 Comparison with ANDI . . . 262
B.4.1 Modified ANDI . . . 264
B.5 Random Splitting Versus Temporal Ordering . . . 268
B.6 Additional Results . . . 270

C Supplement to Chapter 4 **273**
C.1 Asymptotics for Superiority-by-Margin Testing . . . 273
C.2 Parameter Estimation for the SIRDC-SD Model . . . 274
C.3 SIRDCV Model . . . 275
C.4 Evolution of the Epidemic . . . 276
C.4.1 Status Quo . . . 276
C.4.2 Ramp . . . 276
C.4.3 Behavioral . . . 276
C.5 Financial Costs of Vaccine Efficacy Studies . . . 277
C.6 Additional Results . . . 278
C.7 Steps in HCT Setup . . . 312

D Supplement to Chapter 5 **315**
D.1 Company Screening . . . 315
D.2 Orange Book Citations . . . 316
D.3 Initial Public Offerings . . . 318
D.4 Mergers and Acquisitions . . . 320
D.5 Research and Development Pipeline . . . 321
D.6 Drug Approvals . . . 325
D.7 Intellectual Property . . . 329
D.8 FDA Approvals . . . 333
D.9 Additional Discussion . . . 334

E Supplement to Chapter 6 **337**

E.1 S&P Historical Default Rates 337

F Supplement to Chapter 7 **339**

F.1 Literature Estimates 339

F.2 Estimates by Experts 340

F.3 Cost and Duration of GBM AGILE 341

F.4 Correlation 343

F.5 Profitability of a Successful Compound 344

List of Figures

2-1 Sample size of the dataset after filtering 38
2-2 Waterfall plots of DPR 42
2-3 Distributions of key tumor size data 45
2-4 Kaplan-Meier survival curves 47
2-5 Calibration plots of the PFS Cox proportional hazards model 61
2-6 Calibration plots of the OS Cox proportional hazards model 62
2-7 Risk stratification by survival models 63

3-1 Predictive models for regulatory approval 75
3-2 Success rates over time in datasets 79
3-3 Missingness in drug features 81
3-4 Missingness in trial features 82
3-5 Imputation and machine learning methodology 83
3-6 Feature matrix 91
3-7 Comparison of 5NN-RF and ANDI 93
3-8 Distributions of prediction scores 101
3-9 Time-series walk-forward analysis 105
3-10 Time-series walk-forward results 105
3-11 Distributions of prediction scores 107
3-12 Network graph of pipeline predictions for P2APP 108

4-1 Simulation framework 120
4-2 Infections as time-to-event data 130
4-3 Dates of licensure under different clinical trial designs 139

5-1 Highest stage of development of pipeline candidates of MIT licensees 160
5-2 Indication groups of pipeline candidates of MIT licensees 160
5-3 Summary statistics of approved drugs with MIT licensee contribution 164

6-1 Simulation framework for the megafund 176
6-2 Drug development process as a multi-state Markov chain 180
6-3 Distribution functions of successes 186
6-4 Distributions of cumulative ROE 190
6-5 Sensitivity of cumulative ROE 200
6-6 Sensitivity of annualized ROE 201

7-1 Simulation framework for the megafund 209
7-2 Possible development paths for the portfolio 215
7-3 Correlation matrix of brain cancer projects 217
7-4 Investment timeline of a brain cancer drug 219

A-1 Case studies of longitudinal SLD data 239

B-1 Data cleaning for P2APP 244
B-2 Data cleaning for P3APP 245
B-3 Multiple imputation 248
B-4 Datasets created in experiment 255
B-5 Distribution of accrual after imputation 260
B-6 ROC curve of ANDI 265

D-1 Simple moving average of MIT biotech company IPOs 318
D-2 BRDPI adjusted net proceeds for IPOs 319
D-3 IPO dilution 319
D-4 Acquisition values of MIT biotech companies 320
D-5 Cumulative patents granted to MIT licensees 330
D-6 Correlation between drug approvals and patents granted 332

List of Tables

2.1 Characteristics of clinical trials in the dataset 36
2.2 List of variables extracted from SDTM and ADaM databases 37
2.3 Summary statistics of categorical features 39
2.4 Summary statistics of continuous features 39
2.5 Summary statistics of best overall response 41
2.6 Median PFS and OS 46
2.7 Performance of predictive models for tumor response 56
2.8 Top 20 coefficients of the general statistical model 58
2.9 Top 20 coefficients of the general logistic regression model 59
2.10 Performance of predictive models for PFS 60
2.11 Performance of predictive models for OS 60
2.12 Top 20 coefficients of the PFS Cox proportional hazards model 64
2.13 Top 20 coefficients of the OS Cox proportional hazards model 65

3.1 Description of drug and trial features 77
3.2 Sample sizes of datasets 78
3.3 Breakdown by indication groups 78
3.4 Breakdown by sponsor types 79
3.5 Missingness in drug features 82
3.6 Missingness in trial features 83
3.7 Sample size of gold-standard dataset 93
3.8 Out-of-sample performance of missing data approaches 94
3.9 Out-of-sample performance of classifiers 99

3.10 Distributions of prediction scores . . . 100
3.11 Distributions of prediction scores . . . 100
3.12 Top ten most important variables . . . 102
3.13 Distributions of prediction scores . . . 106
3.14 Distribution of prediction scores . . . 106
3.15 Top ten most important variables . . . 109
3.16 Out-of-sample and out-of-time performance for P2APP . . . 110
3.17 Out-of-sample and out-of-time performance for P3APP . . . 111
3.18 Top five pipeline predictions for P2APP by indication groups . . . 112

4.1 Assumptions common across all clinical trial designs . . . 124
4.2 Assumptions specific to each clinical trial design . . . 125
4.3 Sensitivity analysis . . . 136
4.4 Expected number of incremental infections and deaths avoided . . . 137

5.1 Top 30 worldwide top-selling drugs in 2000 . . . 150
5.2 Top 30 worldwide top-selling drugs in 2015 . . . 151
5.3 Summary of MIT portfolio of life science and therapeutics companies 152
5.4 List of MIT IP citations in the Orange Book . . . 154
5.5 IPO data of publicly traded therapeutics companies . . . 156
5.6 Acquisition values of MIT biotech companies . . . 158
5.7 Summary statistics of drug approvals by MIT licensees . . . 163

6.1 Parameters used to simulate a megafund for rare diseases . . . 180
6.2 Distribution functions of successes . . . 185
6.3 Performance of RBO structures . . . 191
6.4 Sensitivity of the vanilla RBO performance . . . 197
6.5 Sensitivity of the guarantee-backed RBO performance . . . 198
6.6 Sensitivity of the equity-only RBO performance . . . 199

7.1 Hypothetical portfolio of brain cancer therapeutics . . . 211
7.2 Parameters for standard clinical trials . . . 213

7.3 Parameters for GBM AGILE 215
7.4 Performance of the NBTS portfolio 221

A.1 List of predictive factors for tumor response 241

B.1 Data pre-processing procedures 245
B.2 Examples of drug and trial features 246
B.3 Biasness of imputations 256
B.4 Out-of-sample performance of missing data approaches 257
B.5 Missingness in drug features 261
B.6 Missingness in trial features 261
B.7 Sample size of oncology-only dataset 264
B.8 Modified ANDI rubric 264
B.9 Oncology ANDI algorithm 267
B.10 Comparison of random splitting and temporal ordering 269
B.11 Out-of-sample and out-of-time performance for P2APP 270
B.12 Out-of-sample and out-of-time performance for P3APP 272

C.1 Expected number of incremental infections and deaths avoided 278
C.2 Expected number of incremental infections and deaths avoided 280
C.3 Expected number of incremental infections and deaths avoided 282
C.4 Expected number of incremental infections and deaths avoided 284
C.5 Expected number of incremental infections and deaths avoided 286
C.6 Expected number of incremental infections and deaths avoided 288
C.7 Expected number of incremental infections and deaths avoided 290
C.8 Expected number of incremental infections and deaths avoided 292
C.9 Estimated date of licensure and probability of approval 294
C.10 Estimated date of licensure and probability of approval 296
C.11 Estimated date of licensure and probability of approval 298
C.12 Estimated date of licensure and probability of approval 300
C.13 Estimated date of licensure and probability of approval 302

C.14 Estimated date of licensure and probability of approval 304
C.15 Estimated date of licensure and probability of approval 306
C.16 Estimated date of licensure and probability of approval 308
C.17 Estimated date of licensure and probability of approval 310

D.1 Expanded list of MIT IP citations in the Orange Book 317
D.2 Pipeline candidates by highest development stage 323
D.3 Pipeline candidates by indication group 324
D.4 List of drugs with MIT licensee contribution 326
D.5 Label expansions by MIT licensees 327
D.6 Post-acquisition drug approvals by MIT licensees 328
D.7 Patents licensed by and granted to MIT licensees 331
D.8 FDA NDA/BLA and NME/ NBE drug approvals 333

E.1 Global corporate average cumulative default rates 338

F.1 Literature estimates of parameters for standard clinical trials 339
F.2 Estimates of parameters for standard clinical trials by NBTS experts 340
F.3 Assumptions for profitability of an approved drug for GBM 344
F.4 NPV of projects on approval 345

Part I

Introduction

Chapter 1

Challenges of Biomedical Innovation

1.1 Introduction

The past two decades have seen an onslaught of biomedical innovations that have revolutionized drug discovery and disease treatment, including gene therapies for diseases thought to be incurable, immunotherapies for cancers, the use of human genome sequencing to discover new treatment modalities, 3D printing of complex biomaterials, and not to mention advances in medical imaging, bioinformatics, and diagnostics. Despite the many promising breakthroughs, studies show that the research and development (R&D) productivity of the pharmaceutical industry has actually been steadily declining since 2000 and is at its lowest level today [1, 2, 3]. Even the most optimistic estimate puts the current pharmaceutical R&D efficiency at levels no higher than twenty years ago [4].

The truth is that biotechnology and pharmaceutical R&D has become more challenging for various reasons. Advances in molecular biology have led to a proliferation of plausible targets to pursue for therapeutic intervention [5, 6]. Most of these genomic targets are highly novel yet poorly validated, making projects based on such targets much riskier undertakings than the well-characterized targets that were developed in the 1990s [7]. The "omics" revolution has also catalyzed a shift in the drug indus-

try away from the "one-size-fits-all" paradigm to personalized medicine approaches optimized based on specific patient characteristics and biomarkers. While highly specific therapeutics show great potential, they are more expensive and time-consuming to develop. More importantly, because such specialized products target only small populations of patients, they generate significantly less revenue as compared to blockbuster drugs a decade ago, especially in the current climate where the pricing of new therapies has come under increasing scrutiny and pressure from regulators, payers, and patients.

Combinatorial drug discovery—an important alternative to the conventional single-agent approach for identifying effective combination therapies to treat complex diseases such cancers and neurological disorders—has also become less efficient over time as the combinatorial chemical search space increases exponentially with each new drug approval. The amount of resources required to search through the sheer number of possibilities has substantially slowed down discovery efforts. Furthermore, an ever-improving back catalog of approved medicines has raised the evidential hurdle for approval, making it increasingly difficult to achieve incremental improvement over time [1]. This crowds R&D activity into hard-to-treat diseases and complex treatment modalities that are potentially more transformative but also riskier to develop. Regulatory hurdles also appear to be rising as we observe a progressive lowering of risk tolerance by regulators [1]. The ever-growing number of safety requirements imposed by the U.S. Food and Drug Administration (FDA) has only made it more costly for biopharma companies to navigate the drug development process [8].

Today, there is significant uncertainty surrounding the scientific, medical, economic, regulatory, academic, and political environments within the biomedical ecosystem. Rising costs of clinical trials, a shift in research focus to more complex scientific pathways that have higher risks of failure, a tougher regulatory environment, tightening of drug pricing legislation, increasing competition from generics, the looming patent cliff for biologics, mounting competitive pressure in emerging markets, and continuing downward pressure by funding organizations on R&D budgets have created a volatile, unconducive environment for investments [2]. These factors have led

to diminishing returns in pharmaceutical R&D, further driving investors away from the biomedical industry to other sectors that can provide more attractive opportunities. According to the National Venture Capital Association (NVCA), the dollar volume of venture capital (VC) investments in the life sciences as proportion of total VC activity in the U.S. was 16.8% in 2019, one of the lowest levels since 2004 when it was as high as 27.7% [9, 10]. The total number of active biotech VC firms and number of biotech initial public offerings in the U.S. and Europe have also declined since 2014, indicating weakened interest from both private and public equity, the traditional funding sources of biotech startups [11, 12]. This outflow of capital has created an ever-widening gap in funding between early-stage basic biomedical research (usually funded by research grants from government agencies such as the National Institutes of Health) and late-stage clinical development (typically financed by large pharmaceutical companies), where many promising academic discoveries go to die not because of bad science but due to financial reasons. This vacuum in the funding of translational R&D is well known in the drug industry as the valley of death.

1.2 Thesis Contributions

In this thesis, we explore the use of data analytics to facilitate biomedical innovation in different areas, with a particular emphasis on the mismatch between the risk characteristics of biomedical projects and the risk preferences of biopharma investors as outlined in Section 1.1. Apart from the introduction and the conclusion, the thesis consists of six chapters, which can be broadly categorized into two themes: clinical trial analytics and new business models. Work on this thesis has led to multiple publications on related topics [13, 14, 15, 16, 17, 18] and several papers that are currently pending submission [19] or under review as of writing [20, 21].

1.2.1 Clinical Trial Analytics

In Chapter 2, we develop data analytics for precision medicine, an emerging area of innovation in disease treatment with the advent of human genome sequencing. While

the prediction of clinical outcomes is central to personalized medicine and the design of clinical trials, especially for a heterogeneous disease like non-small cell lung cancer (NSCLC), there are no predictive models for NSCLC that are widely implemented in practice. In this chapter, we apply survival analysis and machine learning techniques on patient-level clinical trial data to develop prognostic models for response and survival in patients with advanced NSCLC. Our models reflect recent advances in the treatment paradigm of NSCLC, including biomarker-driven personalized treatments such as targeted therapies (e.g., epidermal growth factor receptor tyrosine kinase inhibitors) and immunotherapies (e.g., programmed death-ligand 1 immune checkpoint inhibitors).

In Chapter 3, we turn our attention to the development of better analytics for quantifying and characterizing the risks and uncertainty in biomedical projects. In particular, the probability of success (PoS) of clinical trials is a key parameter that many clinical researchers and biopharma investors consider when making important scientific and business decisions. Without up-to-date estimates, investors may misjudge the risk and value of projects, leading to lost opportunities for both investors and patients. Therefore, having accurate estimates of the PoS is critical for efficient risk management and resource allocation. In this chapter, we apply statistical imputation methods and machine learning algorithms on two large pharmaceutical pipeline databases to develop predictive models for estimating the PoS of drug development programs. The use of artificial intelligence in drug development is not a new concept. Drug developers have already applied machine-learning tools to the discovery process via high-throughput screening of vast libraries of chemical and biological compounds to identify drug targets. However, in managing their portfolios of investigational drugs, biopharma companies typically use unconditional estimates of regulatory approval rates based on historically observed relative frequencies. We propose the use of a wide range of drug and clinical-trial features to obtain conditional estimates of success, and show that our approach achieves promising levels of predictive power. By providing more accurate forecasts of drug development outcomes, and consequently greater risk transparency, our models can help facilitate more accurate

matching of investor risk preferences with the risks of biomedical investment opportunities. Such predictive analytics also reduces the uncertainty surrounding drug development, which will in turn increase the amount of capital that investors are willing to allocate to biomedical projects. By extension, this would lower the cost of capital and increase the efficiency of capital allocation and portfolio decision-making.

In Chapter 4, we focus on analytics related to the ongoing coronavirus pandemic. The world is facing unprecedented challenges from the COVID-19 (coronavirus disease 2019) pandemic. Given the dire situation, human challenge clinical trials (HCTs) have been proposed as a way to expedite the vaccine development process. While moral concerns have been raised, bioethicists generally agree that an HCT may be ethically permissible if it can provide greater societal value versus traditional pathways. However, there has not been any quantitative analysis of the potential benefits of a COVID-19 HCT versus non-challenge trials in literature, thus making it difficult to justify the use of a challenge study at this time. In this chapter, we propose a systematic, transparent, reproducible, and principled simulation framework for quantitatively assessing the potential costs and benefits of different vaccine efficacy clinical trial designs for COVID-19 vaccine development, including traditional and adaptive randomized clinical trials, and HCTs. Our results contribute to the moral and ethical debate about HCTs by identifying situations where HCTs can provide greater social value versus conventional development pathways, and are thus justifiable. Our methodology allows stakeholders, such as vaccine developers, policymakers, and HCT volunteers to understand the implications of their actions (or inaction), and to make more informed ethical decisions regarding accelerating COVID-19 vaccine development amidst this crisis.

1.2.2 New Business Models

In Chapter 5, we perform a systematic study of technology licensing by the Massachusetts Institute of Technology (MIT) in the therapeutics domain. The process of drug development in the pharmaceutical industry is undergoing a profound shift in its industrial organization. Instead of relying on in-house research, big pharmaceutical

companies are deploying growing amounts of capital previously committed to internal R&D to acquire late stage, de-risked clinical assets with nearer-term payoffs to replenish their development pipelines [22]. On the other hand, smaller biotechnology firms have taken a more active role in early-stage drug discovery. Academic institutions also play an increasingly critical role in the industry through the licensing of seminal discoveries and the creation of startups. Despite the growing importance of technology licensing to the biomedical ecosystem, there has been surprisingly little data collected on the impact of technology transfer by academia. In this chapter, we address the knowledge and data gap through a systematic analysis of the financial and research impact of MIT life sciences technology licensing. We construct several measures of impact including MIT patents cited in the Orange Book, capital raised, outcomes from mergers and acquisitions, patents granted to MIT intellectual property licensees, drug candidates discovered, and U.S. drug approvals, a key benchmark of innovation in the biopharma industry. Our methodology provides a useful framework for other academic institutions to track the outcomes of their intellectual property in the therapeutics domain.

As discussed in Section 1.1, traditional sources of financing in biopharma R&D, such as private and public equity, and VC, are no longer effective nor adequate for supporting early-stage translational research, which corresponds to the riskiest and most challenging part of the biomedical innovation process. Due to increasing complexity and risk, the needs and expectations of limited partners and shareholders have become less aligned with the new realities of biomedical innovation. For example, the constant scrutiny of corporate performance has steered the senior management of public companies towards projects with surer and nearer-term payoffs, and away from more speculative but potentially transformative research [23]. According to the Dow Jones VentureSource, less than 4% of the biotech companies funded by VCs in 2014 were in the seed stage [24]. In contrast, almost 80% of the biotech companies that received VC investments were already in product development, indicating the lack of interest and support from VCs in early-stage startups. This is not surprising, given that drug development is widely accepted as one of the most complex and riskiest

businesses that is not only subjected to scientific challenges but also vulnerable to external economic and public policy conditions.

Biomedical projects are difficult to fund on a standalone basis because they require a large amount of initial capital, have long gestation lags during which no cash flows are generated and additional investments are needed, and perhaps most importantly, have low probabilities of success [25]. The average drug requires at least a decade of translational research and clinical testing before it is approved by the FDA. Because of these characteristics, the funding requirements of biomedical research far outstrip the capital available from traditional sources of funding, thus creating a valley of death. In Chapters 6 and 7, we study the use of financial engineering techniques such as portfolio theory and securitization to mitigate and structure the risks inherent in biomedical projects. In particular, we consider the recently proposed megafund approach [23], which involves combining a large number of biomedical assets into a single portfolio to diversify the financial risk of therapeutic development and increase the likelihood of success through multiple "shots on goal." Although it is impossible for any VC to fund a portfolio of such scale (requiring capital between hundreds of millions to several billion dollars to achieve sufficient risk reduction) singlehandedly, the megafund can tranched—that is, securitized—to create equity and investment grade bonds with risk-reward characteristics that are attractive to institutional investors. This allows the fund to tap into the fixed income market, a substantially larger pool of capital than the conventional sources of biopharma R&D financing but one traditionally unwilling to participate in biopharma investments due to the risky and fragmented nature of drug development. According to the Securities Industry and Financial Markets Association, the size of the U.S. bond market was $45 trillion in 2019, which is two orders of magnitude larger than the $444 billion in assets under management by VCs in the same year, as reported by the NVCA [10, 26].

In Chapter 6, we extend the recently proposed megafund structure to account for technical correlation between assets in the underlying portfolio using a single-factor model with a Gaussian copula, thus making it a more realistic representation of biopharma R&D, and also allowing us to evaluate the tail risks of the megafund more

accurately—the financial crisis of 2008 has made clear the importance of correlations between underlying assets in the valuation of asset-backed securities [27]. In Chapter 7, we further extend the megafund framework to include adaptive clinical trial designs, and demonstrate the economic viability of using the megafund vehicle to finance and accelerate drug development for glioblastoma, a disease with very few treatment options, low probabilities of success, and huge unmet need.

Part II

Clinical Trial Analytics

Chapter 2

Predictive Models for Patient Outcomes in Lung Cancer

Lung cancer is the leading cause of cancer-related mortality in the world. In particular, non-small cell lung cancer (NSCLC) accounts for approximately 85% of lung cancer cases. Recent advances in molecularly targeted therapy and immunotherapy have changed the treatment paradigm of NSCLC. An updated predictive model for clinical outcomes that reflects the current standard of care for advanced-stage NSCLC has broad clinical utility in terms of developing individualized treatment plans and risk stratification. In this chapter, we aggregate data from 17 randomized clinical trials submitted to the U.S. Food and Drug Administration, evaluating chemotherapy, targeted therapy, and immunotherapy in patients with advanced NSCLC. We develop and validate a range of statistical and machine-learning predictive models for three important clinical endpoints—objective response (OR), progression-free survival (PFS) and overall survival (OS)—in NSCLC patients using routinely collected patient and disease variables, including biomarker mutations, and inhibitor therapy. Our models achieved promising out-of-sample predictive performances. We find biomarker status to be the strongest predictor of OR, PFS, and OS in patients treated with immune checkpoint inhibitors and targeted therapies. However, single biomarkers have limited predictive value, especially for immunotherapy. To advance beyond the results achieved in this study, data on composite multi-omic signatures is required.

2.1 Introduction

Lung cancer is one of the most commonly diagnosed cancers in the United States and worldwide, and the leading cause of cancer-related mortality. In particular, non-small cell lung cancer (NSCLC) accounts for approximately 85% of lung cancer cases. The majority of NSCLC patients are diagnosed at advanced stages (III and above) [28]. The standard of care for these patients is typically chemotherapy. However, recent advances in molecularly targeted therapy and immunotherapy have been shown to significantly improve the survival of specific patient groups.

Predictive models play an important role in cancer treatment planning. These models, by providing accurate predictions of the survival rate, allow patients to make more informed decisions about treatment. Because cases of NSCLC comprise a heterogeneous group of patients, there is a wide variation in the effectiveness of different therapies. With predictive models for clinical outcomes, physicians can develop treatment plans based on the specific characteristics of individual patients rather than on general statistics of the population. In addition, predictive models can be used to support patient selection and risk stratification in clinical trials. Despite their clinical relevance, however, there are no predictive models for NSCLC that are widely implemented in practice. Mahar et al. [29] reviewed 32 lung cancer prognostic tools published between 1996 and 2015. They found many studies to be poorly designed and inadequately described. Most did not conduct a formal evaluation of the internal validity of the developed model. Some contained novel but expensive and difficult to measure factors that would be impractical to include in prognostic models intended for common clinical use.

In this chapter, we perform a pooled analysis of 17 randomized clinical trials in NSCLC submitted to the U.S. Food and Drug Administration (FDA) to support New Drug Applications. The trials evaluated chemotherapy, targeted therapy, and immunotherapy treatments in patients with advanced NSCLC. We characterize the tumor dynamics, response, progression-free survival (PFS), and overall survival (OS) of patients under these different treatment modalities. Our aim is to develop updated

predictive models for three important clinical endpoints—tumor response, PFS, and OS—that reflect recent advances in the treatment paradigm of NSCLC. To this end, we propose a stochastic tumor growth model based on the longitudinal tumor size data collected in clinical trials to predict tumor response. At the same time, we explore machine-learning algorithms and survival models. In our models, we consider clinical, demographic, and pathological features routinely collected in medical screenings. We describe our training and testing methodology in Section 2.3. At the end, we identify baseline variables that are strongly associated with response and survival, and compare our findings with related studies in the literature.

2.2 Data

2.2.1 Study Population

We specify 17 randomized clinical trials submitted to the FDA between January 2007 and February 2017 as our initial dataset. These trials evaluate treatments under nine approved drugs for NSCLC, consisting of three programmed death-ligand 1 (PDL1) immune checkpoint inhibitors (ICI), three epidermal growth factor receptor (EGFR) tyrosine kinase inhibitors (TKI), and three anaplastic lymphoma kinase (ALK) TKIs. We summarize the characteristics of the clinical trials—experimental and control arms, trial design, line of therapy, and sample size—in Table 2.1. Five trials test immunotherapy, one in the first-line setting and four in the second-line setting; four trials assess ALK-translocation targeted therapy, two in the first-line setting and two in the second-line setting; and eight trials evaluate EGFR-mutation targeted therapy, four in the first-line setting, three in the second-line setting, and one in the third-line setting. Almost all trials are open-label with standard-of-care chemotherapy as the control arm. In aggregate, the dataset includes 8,925 patients in the intention-to-treat population.

We extract survival data, tumor measurements, response outcomes, baseline demographics, medical history, and laboratory tests results from patient-level Study

Data Tabulation Model (SDTM) and Analysis Data Model (ADaM) databases. After compiling and standardizing common features across all trials, we have 46 categorical variables and 5 continuous variables in the dataset (excluding endpoints, see Table 2.2).

All trials in our dataset adopt the Response Evaluation Criteria in Solid Tumors (RECIST) for tumor measurements and response outcomes. In cases where both investigator-assessed and independent review committee (IRC)-determined tumor measurements are available, we give priority to the version used by the investigators to decide the continuation or discontinuation of the study medication, as stated in the clinical trial protocol. For response outcomes, we give priority to confirmed outcomes over unconfirmed ones, and use values assessed by IRC whenever available, since they are generally perceived to be more reliable and less susceptible to bias compared to investigator-assessed outcomes—e.g., IRC-confirmed best overall responses are often used to compute the objective response rate as a secondary endpoint in clinical trials. We do the same for PFS. Five studies in the dataset were initiated under the older RECIST version 1.0, and use the sum of longest diameter (SLD) measurements based on up to ten measurable target lesions. This limit was reduced to five lesions in RECIST version 1.1. In order to reconcile measurements collected under the older criteria with the current version, we scale earlier measurements to reflect the new five lesion limit (see Appendix A.1).

For our analysis, we exclude patients who either (1) did not have tumor measurements in the database, or had ambiguous records, such as non-measurable disease, no target lesions, or a 0 mm baseline SLD, (2) were given a placebo (e.g., the placebo comparator arm in the third-line Afatinib trial) or were not otherwise treated by chemotherapy, immunotherapy, or targeted therapy in the clinical trial before its discontinuation, or (3) had missing features in their records that were necessary for subsequent analyses. The final sample comprises 7,805 patients (see Fig. 2-1). In Tables 2.3 and 2.4, we pool patients by the type of therapy received—chemotherapy, PDL1 ICI, EGFR TKI, and ALK TKI—and list the summary statistics of key baseline demographics and medical history to give the reader an intuition for the characteris-

tics of the dataset.

Most patients in our dataset have advanced NSCLC with some form of metastasis. All patients treated with ALK TKI are proven positive for the ALK mutation. In contrast, about two-thirds of the patients (67%) under EGFR TKI have an unknown EGFR mutation status. The proportion of patients in the sample with ALK rearrangements (9%) is almost twice that observed in the general NSCLC population (5%) [30]. The overall median patient age is 60 years, with that for the ALK trials being lower at 53 years. Unlike other therapy groups, about 62% of the ALK sample have no history of smoking, and over 90% are diagnosed with adenocarcinoma. This is consistent with studies showing that ALK translocations are observed predominantly in adenocarcinomas, and among younger and nonsmoking patients [31].

Most of the patients are enrolled outside the United States, mainly in the Asia-Pacific and the Western Europe regions. Over half of the patients in the dataset are white (58%). In particular, PDL1 ICI seems to be much more well-studied in Caucasians (80%) than Asians (15%). In general, there is an even mix of both sexes in the dataset, except in the PDL1 group, where over 60% of the patients are male. Since more than half of the trials in the dataset are in the second-line setting or higher, the majority of the patients (65%) have undergone at least one regimen of chemotherapy prior to participation in these clinical trials.

2.2.2 Tumor Response Data

Tumor response is an important efficacy endpoint in cancer clinical trials, and one of the most commonly used. The use of tumor regression for evaluating cancer therapeutics is supported by multiple studies that demonstrate an association between solid tumor shrinkage and improved OS, or to other time-to-event measures, such as PFS [32]. It is typically employed as a secondary endpoint to complement survival data, but tumor response has been used as the primary surrogate endpoint in some single-arm trials to support the accelerated approval of breakthrough therapies and orphan drugs, together with the duration of response. An example is Osimertinib, which received accelerated approval in November 2015 based on an objective response

Table 2.1: Characteristics of clinical trials in the dataset. *Abbreviations*: Chemo, chemotherapy; R, randomized; OL, open-label; DB, double-blind; ITT, intention-to-treat population.

Therapy	**Treatment**	**Therapy**	**Control**	**Design**	**Phase**	**Line**	**ITT**	**Initiation**	**Cutoff**
EGFR	Gefitinib	Chemo	Docetaxel	R, OL	3	2nd	1,466	Mar-04	Mar-07
EGFR	Gefitinib	Chemo	Carboplatin with paclitaxel	R, OL	3	1st	1,217	Mar-06	Apr-08
EGFR	Erlotinib	Chemo	Pemetrexed or docetaxel	R, OL	3	2nd	424	Apr-06	Aug-10
EGFR	Erlotinib	Chemo	Docetaxel or gemcitabine with cisplatin or carboplatin	R, OL	3	1st	173	Feb-07	Apr-12
EGFR	Afatinib	Placebo	Best supportive care	R, DB	2/3	3rd	585	Apr-08	Jun-10
EGFR	Afatinib	Chemo	Pemetrexed with cisplatin	R, OL	3	1st	345	Aug-09	Nov-13
ALK	Crizotinib	Chemo	Pemetrexed or docetaxel	R, OL	3	2nd	347	Sep-09	Aug-15
ALK	Crizotinib	Chemo	Pemetrexed with cisplatin or carboplatin	R, OL	3	1st	343	Jan-11	Nov-13
EGFR	Erlotinib	Chemo	Gemcitabine with cisplatin	R, OL	3	1st	217	Mar-11	Apr-14
EGFR	Afatinib	EGFR	Erlotinib	R, OL	3	2nd	795	Mar-12	Feb-15
PDL1	Nivolumab	Chemo	Docetaxel	R, OL	3	2nd	272	Oct-12	Dec-14
PDL1	Nivolumab	Chemo	Docetaxel	R, OL	3	2nd	582	Nov-12	Feb-15
ALK	Ceritinib	Chemo	Pemetrexed or docetaxel	R, OL	3	2nd	231	Jun-13	Jan-16
PDL1	Atezolizumab	Chemo	Docetaxel	R, OL	2	2nd	287	Aug-13	Dec-15
PDL1	Pembrolizumab	Chemo	Docetaxel	R, OL	2/3	2nd	1,033	Aug-13	Oct-15
ALK	Alectinib	ALK	Crizotinib	R, OL	3	1st	303	Aug-14	Feb-17
PDL1	Pembrolizumab	Chemo	Platinum-based	R, OL	3	1st	305	Sep-14	May-16
Total							**8,925**		

Table 2.2: List of variables extracted from SDTM and ADaM databases. *Abbreviations*: APAC, Asia-Pacific; NAM, North America; WEUR, Western Europe; Adeno, adenocarcinoma; SCC, squamous cell carcinoma; CR, complete response; PR, partial response; SD, stable disease; PD:, progressive disease; NE, not evaluable.

Type	**Variable**	**Values**
Demographics	Age	Years
	Weight	kg
	Sex	Male, female
	Race group	Asian[1], white, others
	Region	APAC, NAM, WEUR, others
Medical history	Time since diagnosis	Days
	Performance status[2]	0, 1, 2 or higher
	Smoking status	Ever, never
	Stage at screening	IIIB or lower, IV
	Prior chemotherapy	Yes, no
	Histology	Adeno, SCC, others[3]
	Metastases in brain, bone, liver and others	Yes, no
	Number of metastasis sites	Count
	Biomarker status in PDL1, EGFR and ALK[4]	Positive, negative, not tested
	Number of baseline target lesions	1, 2, 3, 4, 5 or more
	Baseline SLD[5]	mm
	Comorbidities in 23 system organ class levels[6]	Yes, no
Laboratory measurements[7]	Alkaline phosphate	High, normal, low
	Alanine aminotransferase	High, normal, low
	Aspartate aminotransferase	High, normal, low
	Bilirubin	High, normal, low
	Creatine	High, normal, low
	Hemoglobin	High, normal, low
	Platelets count	High, normal, low
	White blood cells count	High, normal, low
Therapy type	Therapy received	Chemotherapy, PDL1 ICI, EGFR TKI, ALK TKI
Endpoints	Overall survival	Days
	Overall survival censor	Yes, no
	Progression-free survival	Days
	Progression-free survival censor	Yes, no
	Best overall response	CR, PR, SD[8], PD, NE
	Objective response[9]	Yes, no
	Timepoint SLD	mm
	Depth of response	%

[1] Includes Pacific Islanders. [2] Eastern Cooperative Oncology Group (ECOG) or World Health Organization (WHO) score. [3] Includes large cell carcinoma (LCC) and not otherwise specified (NOS).
[4] Patients are tested for at most one biomarker depending on the experimental arm of the the clinical trial they are from: patients from PDL1 ICI trials are tested for PDL1 expression, EGFR TKI trials for EGFR-mutation, and ALK TKI trials for ALK-translocation. [5] Measurements under RECIST version 1.0 are scaled to reconcile with version 1.1 (see Appendix A.1). [6] As defined in the Medical Dictionary for Regulatory Activities (MedDRA).
[7] High, normal, low as determined by investigators on-site. [8] Includes non-CR/non-PD.
[9] Defined as having either CR or PR as best overall response.

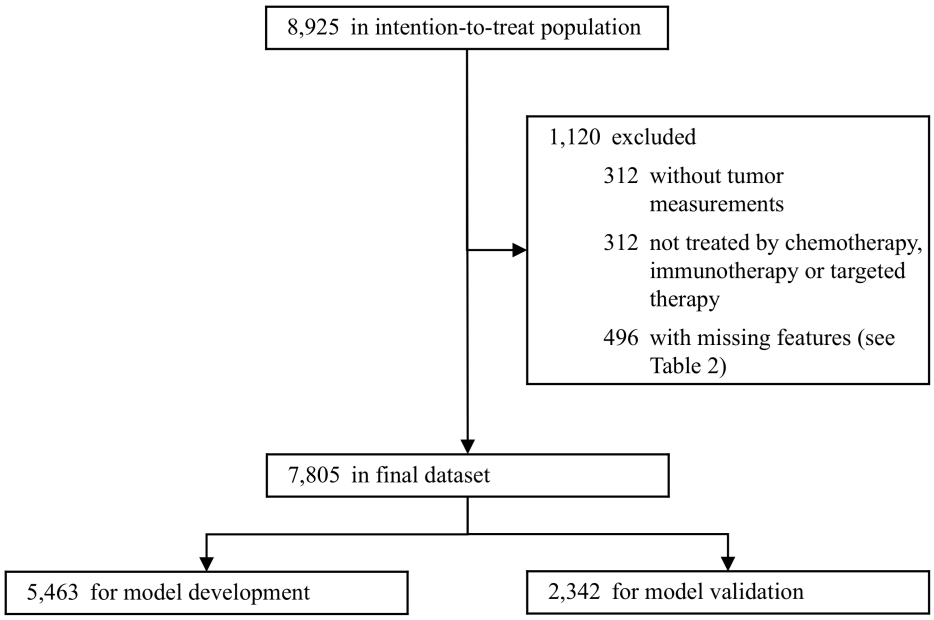


Figure 2-1: Sample size of the dataset after filtering. We exclude patients who either (1) did not have tumor measurements in the database or had ambiguous records such as non-measurable disease, no target lesions, or 0 mm baseline SLD, (2) were on placebo (e.g., the placebo comparator arm in the third-line Afatinib trial), or were not treated by either chemotherapy, immunotherapy, or targeted therapy in the clinical trial before discontinuation, or (3) had missing features that are necessary for subsequent analyses (see Table 2.2). We randomly select 70% of the dataset as the development cohort and use the other 30% as the validation cohort (see Section 2.3.2).

Table 2.3: Summary statistics of key baseline demographics and disease characteristics (categorical features) broken down by the type of therapy received.

		Chemo (n=2,999)		PDL1 ICI (n=1,323)		EGFR TKI (n=2,781)		ALK TKI (n=702)		All (n=7,805)	
Variable	**Group**	**n**	**%**	**n**	**%**	**n**	**%**	**n**	**%**	**n**	**%**
Sex	Male	1,518	50.6	807	61	1,526	54.9	308	43.9	4,159	53.3
	Female	1,481	49.4	516	39	1,255	45.1	394	56.1	3,646	46.7
Race group	Asian	1,215	40.5	204	15.4	1,398	50.3	303	43.2	3,120	40
	White	1,715	57.2	1,061	80.2	1,330	47.8	383	54.6	4,489	57.5
	Others	69	2.3	58	4.4	53	1.9	16	2.3	196	2.5
Region	APAC	1,232	41.1	213	16.1	1,389	49.9	305	43.4	3,139	40.2
	NAM	413	13.8	380	28.7	215	7.7	93	13.2	1,101	14.1
	WEUR	870	29	500	37.8	743	26.7	213	30.3	2,326	29.8
	Others	484	16.1	230	17.4	434	15.6	91	13	1,239	15.9
Performance status	0	957	31.9	421	31.8	770	27.7	359	51.1	2,507	32.1
	1	1,839	61.3	900	68	1,807	65	301	42.9	4,847	62.1
	2 or higher	203	6.8	2	0.2	204	7.3	42	6	451	5.8
Smoking status	Ever	1,703	56.8	1,116	84.4	1,592	57.2	268	38.2	4,679	59.9
	Never	1,296	43.2	207	15.6	1,189	42.8	434	61.8	3,126	40.1
Stage at screening	IIIB or lower	363	12.1	107	8.1	400	14.4	16	2.3	886	11.4
	IV	2,636	87.9	1,216	91.9	2,381	85.6	686	97.7	6,919	88.6
Prior chemotherapy	Yes	1,844	61.5	1,159	87.6	1,812	65.2	258	36.8	5,073	65
	No	1,155	38.5	164	12.4	969	34.8	444	63.2	2,732	35
Histology	Adeno	2,210	73.7	800	60.5	1,662	59.8	660	94	5,332	68.3
	SCC	476	15.9	343	25.9	935	33.6	6	0.9	1,760	22.5
	Others	313	10.4	180	13.6	184	6.6	36	5.1	713	9.1
Metastasis	Brain	313	10.4	150	11.3	178	6.4	254	36.2	895	11.5
	Bone	864	28.8	394	29.8	665	23.9	283	40.3	2,206	28.3
	Liver	610	20.3	287	21.7	460	16.5	187	26.6	1,544	19.8
	Others	2,926	97.6	1,282	96.9	2,559	92	697	99.3	7,464	95.6
PDL1 expression	Positive	431	14.4	591	44.7	0	0.0	0	0.0	1,022	13.1
	Negative	418	13.9	656	49.6	0	0.0	0	0.0	1,074	13.8
	Not tested	2,150	71.7	76	5.7	2,781	100	702	100.0	5,709	73.1
EGFR-mutation	Positive	538	17.9	0	0.0	677	24.3	0	0.0	1,215	15.6
	Negative	223	7.4	0	0.0	233	8.4	0	0.0	456	5.8
	Not tested	2,238	74.6	1,323	100.0	1,871	67.3	702	100.0	6,134	78.6
ALK-translocation	Positive	404	13.5	0	0.0	0	0.0	702	100.0	1,106	14.2
	Negative	1	0.0	0	0.0	0	0.0	0	0.0	1	0.0
	Not tested	2,594	86.5	1,323	100.0	2,781	100	0	0.0	6,698	85.8

Table 2.4: Summary statistics of key baseline demographics and disease characteristics (continuous features) broken down by the type of therapy received.

	Median (min–max)				
Variable	**Chemo**	**PDL1 ICI**	**EGFR TKI**	**ALK TKI**	**All**
Age (years)	60 (19–85)	62 (20–90)	61 (24–88)	53 (18–91)	60 (18–91)
Time since diagnosis (days)	164 (1–7,207)	312 (21–11,068)	183 (1–5,503)	87 (11–4,734)	201 (1–11,068)
Baseline SLD (mm)	70 (9–277)	79 (10–298)	68 (10–760)	59 (10–274)	70 (9–760)

rate (ORR) endpoint for the treatment of patients with metastatic EGFR T790M mutation-positive NSCLC who had progressive disease following first-line EGFR TKI therapy.

We summarize the RECIST best overall response outcomes by the type of therapy received and the biomarker mutation status of the patient in Table 2.5. The ORR of patients under inhibitor therapy with the corresponding biomarker mutation (37% for PDL1 ICI, 45% for EGFR TKI, and 65% for ALK TKI) is more than twice as large as those with the same mutations but under chemotherapy (17%, 23%, and 29%). Patients with a negative biomarker status, but still treated with inhibitor therapy, have the worst outcomes. They exhibit progressive disease (PD) the most frequently (43-44%), and have the lowest ORR, even when compared with their counterparts in the chemotherapy group. On average, patients with an unknown EGFR-mutation status seem to respond better to chemotherapy (ORR 16%) than EGFR TKI (ORR 14%).

In Fig. 2-2, we show how the distributions of depth of response (DPR), defined as the maximum reduction in tumor burden with respect to the baseline SLD, differ among the four treatment groups. In these waterfall plots, we sort the DPR in descending order and plot it from left to right. Positive values represent growth, while negative values correspond to shrinkage. Note that having a DPR smaller than -30% does not always lead to an objective response, because the response may not be confirmed in subsequent readings. It is clear that patients under inhibitor therapy with the corresponding biomarker mutation experience a greater reduction in SLD. This is shown by the shift in the waterfalls to the left compared to subgroups with other mutations. Of the therapies, it seems that ALK TKI has the most aggressive anti-tumor activity. Almost all ALK-positive patients demonstrate some extent of tumor regression, and their ORR is substantially larger (65%) than the comparable values for PDL1 ICI and EGFR TKI.

Table 2.5: Summary statistics of best overall response broken down by the type of therapy received and biomarker mutation status.

		CR		PR		SD		PD		NE		ORR	
	N	**n**	**%**	**n**	**%**	**n**	**%**	**n**	**%**	**n**	**%**	**n**	**%**
Chemotherapy													
All	2,999	4	0.1	560	18.7	1,297	43.2	801	26.7	337	11.2	564	18.8
PDL1-positive	431	1	0.2	74	17.2	177	41.1	127	29.5	52	12.1	75	17.4
PDL1-negative	418	0	0	62	14.8	161	38.5	145	34.7	50	12	62	14.8
EGFR-positive	538	1	0.2	121	22.5	265	49.3	93	17.3	58	10.8	122	22.7
EGFR-negative	223	0	0	30	13.5	114	51.1	53	23.8	26	11.7	30	13.5
ALK-positive	404	2	0.5	115	28.5	138	34.2	122	30.2	27	6.7	117	29
ALK-negative	1	0	0	0	0	0	0	1	100	0	0	0	0
Not tested	984	0	0	158	16.1	442	44.9	260	26.4	124	12.6	158	16.1
PDL1 ICI													
All	1,323	11	0.8	306	23.1	397	30	490	37	119	9	317	24
PDL1-positive	591	9	1.5	211	35.7	159	26.9	167	28.3	45	7.6	220	37.2
PDL1-negative	656	1	0.2	83	12.7	213	32.5	292	44.5	67	10.2	84	12.8
Not tested	76	1	1.3	12	15.8	25	32.9	31	40.8	7	9.2	13	17.1
EGFR TKI													
All	2,781	8	0.3	575	20.7	1,027	36.9	858	30.9	313	11.3	583	21
EGFR-positive	677	5	0.7	300	44.3	216	31.9	112	16.5	44	6.5	305	45.1
EGFR-negative	233	0	0	26	11.2	81	34.8	101	43.3	25	10.7	26	11.2
Not tested	1,871	3	0.2	249	13.3	730	39	645	34.5	244	13	252	13.5
ALK TKI													
All	702	27	3.8	431	61.4	149	21.2	60	8.5	35	5	458	65.2
ALK-positive	702	27	3.8	431	61.4	149	21.2	60	8.5	35	5	458	65.2
ALK-negative	0	0	-	0	-	0	-	0	-	0	-	0	-
Not tested	0	0	-	0	-	0	-	0	-	0	-	0	-
Total	**7,805**	**50**	**0.6**	**1,872**	**24.0**	**2,870**	**36.8**	**2,209**	**28.3**	**804**	**10.3**	**1,922**	**24.6**

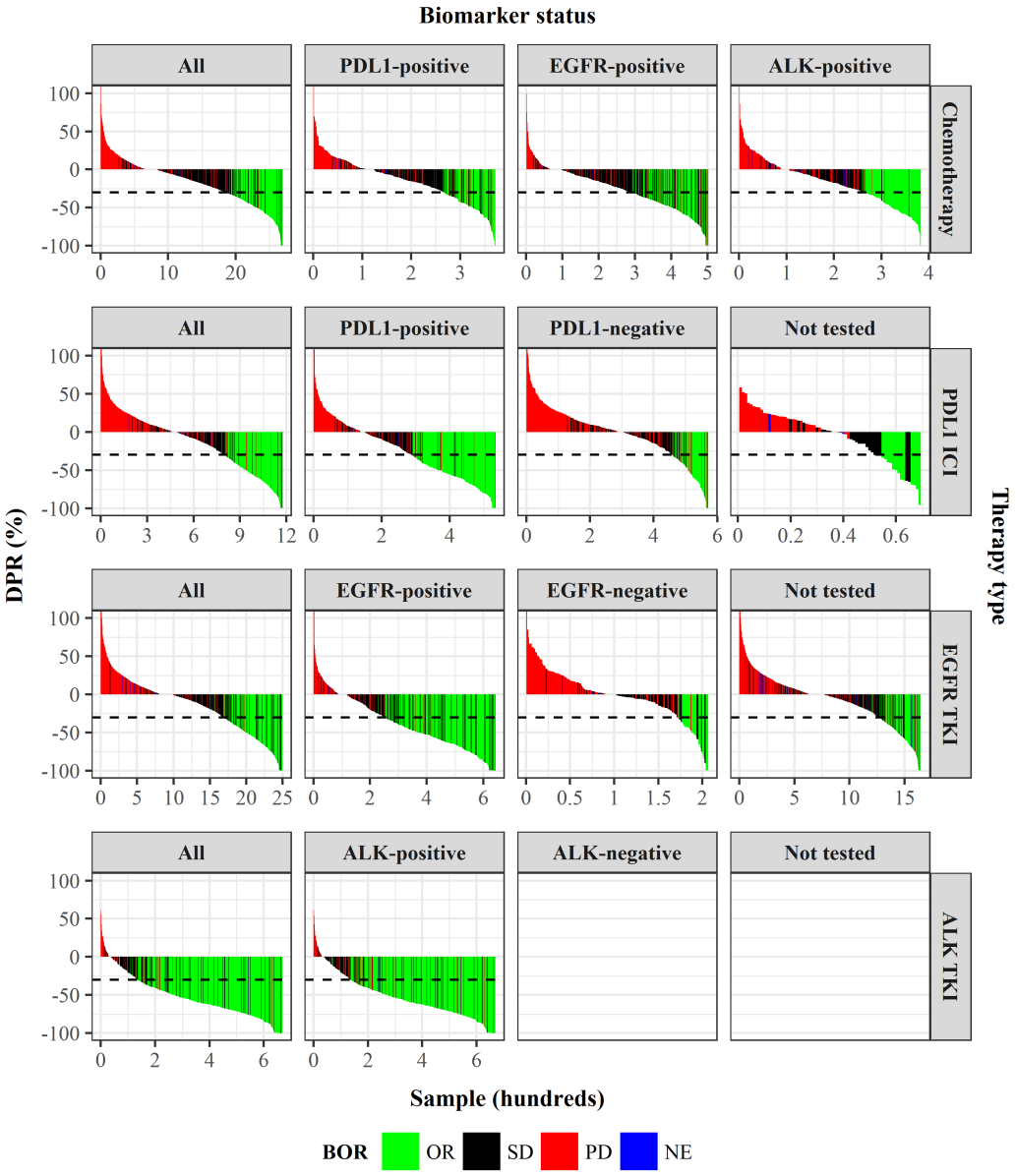


Figure 2-2: Waterfall plots of DPR broken down by the type of therapy received and biomarker mutation status. For example, the second plot on the first row shows the DPR of PDL1-positive patients under chemotherapy and the third plot on the second row shows the DPR of PDL1-negative patients under PDL1 ICI. Each vertical bar represents a single subject. Dashed horizontal line at -30% shows the RECIST threshold for response. We truncate the plots at +100% for better visualization. There are 27 patients who demonstrate tumor growth beyond this limit. We also exclude 774 patients that have only the baseline measurement in the database.

2.2.3 Longitudinal Tumor Size Data

In order to develop a continuous-time and continuous-SLD stochastic tumor growth model to incorporate into our predictive model, we use longitudinal tumor size data extracted from our dataset. This data consists of patient-level SLD measurements collected at clinical trial visits. Studies typically record three types of clinical trial visit, namely the baseline visit, the treatment visits, and the follow-up visits. We use timepoint measurements from all three types of visit in our analysis.

In Fig. 2-3, we show the distributions of key data characteristics broken down by the type of therapy received. In aggregate, the baseline SLDs (BSLDs) seem to follow a shifted log-normal distribution with a mode between 25–35 mm. Patients under chemotherapy and PDL1 ICI generally have larger BSLDs, with modes in the 45–55 mm and 45–65 mm ranges, respectively. In contrast, patients under ALK TKI have the smallest BSLDs on average, with a mode between 15–25 mm. In the dataset, about 10% of the patients completed only one visit, the screening visit. These patients withdrew from the trial after receiving the first dose, but before the first post-baseline tumor assessment visit. While the distributions for chemotherapy, PDL1 ICI and EGFR TKI are largely skewed to the left, that for ALK TKI is quite evenly spread out between 1–12 visits. In general, we rarely observe patients with more than ten visits. Visits are mostly scheduled at intervals of 5–6 weeks, although we observe a second peak for PDLI ICI and ALK TKI at about 9 weeks and 8 weeks, respectively.

We examine several case studies in Appendix A.2 to illustrate some of the subtleties present in the longitudinal SLD dataset. Apart from some straightforward cases, it is often difficult to glean the exact reasons for discontinuations in tumor assessment from the dataset. Nevertheless, it is clear that there is a discontinuation process at work that affects our observation of SLD measurements. For example, we are less likely to observe measurements after PD in SLD has occurred. This phenomenon is in inherent to the data collection process because of the patient safety protections designed into the clinical trials. In this chapter, we will develop a proba-

bilistic model for the discontinuation mechanism, and incorporate it into our tumor growth models.

2.2.4 Survival Data

Overall survival is the gold standard for efficacy endpoints in clinical trials. It is defined as the time from randomization in a clinical trial until death from any cause. While OS is universally recognized as a direct measure of clinical benefit, there are multiple limitations associated with its use. For ethical reasons, many trials allow patients in the control arm who experience progressive disease to receive the experimental therapy (i.e., crossovers). This makes it difficult to assess the impact of the new drug on OS. Moreover, large sample sizes and extended follow-ups are required to demonstrate statistical differences in OS between treatment arms [33]. In many cases, PFS is used as a surrogate endpoint for OS to support drug approvals. Similar to OS, PFS is defined as the time from randomization to tumor progression or death. However, PFS is more easily established because it requires a smaller sample size and a shorter follow-up time. In addition, PFS is free from the confounding effects of crossovers and post-trial therapies.

We extract both survival endpoints from our dataset, and summarize them in Table 2.6 and Fig. 2-4. The median PFS and OS in the dataset are 127 and 384 days, respectively. We observe PFS events in about 76% of the patients and OS events in about 60%. We categorize patients into six treatment groups and examine their survival outcomes, according to the type of therapy received and their biomarker mutation status: (1) any biomarker status and under chemotherapy, (2) PDL1-positive and under PDL1 ICI, (3) EGFR-positive and under EGFR TKI, (4) ALK-positive and under ALK TKI, (5) negative biomarker status but under inhibitor therapy (that is, either PDL1-negative but under PDL1 ICI or EGFR-negative but under EGFR TKI), and (6) not tested for any biomarkers but under inhibitor therapy (either PDL1 ICI or EGFR TKI). In Table 2.6, we find that immunotherapy and targeted therapy offer PFS improvements for patients with positive biomarkers. These patients have a higher median survival compared to other treatment groups. The Kaplan-Meier

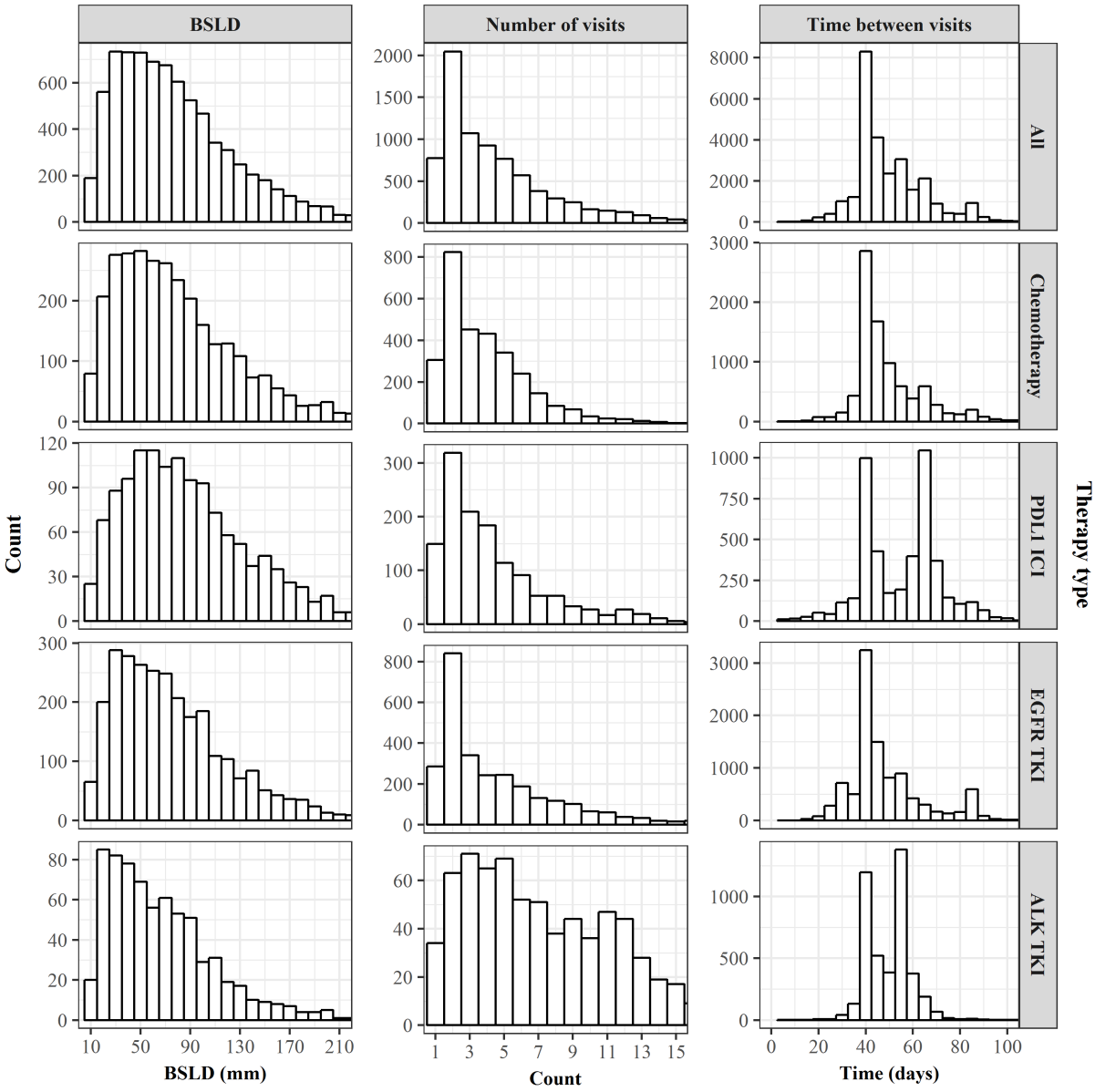
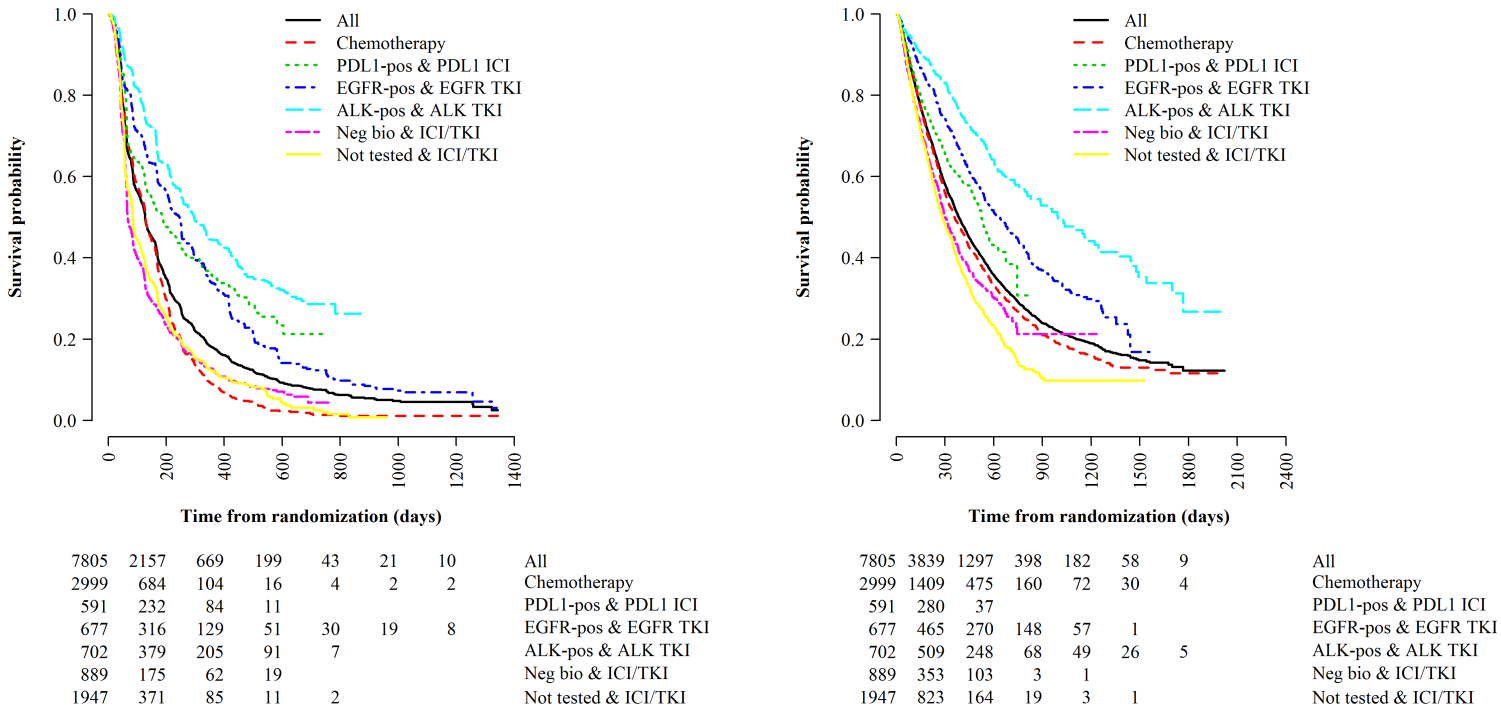


Figure 2-3: Distributions of key tumor size data characteristics broken down by the type of therapy received. The first column shows the BSLDs (scaled as described in Section 2.2), second the number of visits completed by each patient, and third the time intervals between two consecutive visits for all patients in the dataset. We truncate the BSLD at 215 mm, number of visits at 15, and time between visits at 102 days for better visualization. There are 113 patients with BSLD larger than 215 mm, 97 patients who completed more than 15 visits, and 139 pairs of consecutive visits that are apart by more than 102 days.

Table 2.6: Median PFS and OS broken down by treatment group.

		PFS (days)		OS (days)	
Treatment Group	**N**	**Events**	**Median**	**Events**	**Median**
All	7,805	5,959	127	4,712	384
Chemotherapy	2,999	2,406	127	1,884	359
PDL1-positive and under PDL1 ICI	591	369	186	247	522
EGFR-positive and under EGFR TKI	677	488	246	402	639
ALK-positive and under ALK TKI	702	409	295	264	994
Negative biomarker but under inhibitor therapy	889	764	66	544	299
Not tested but under inhibitor therapy	1,947	1,523	84	1,371	283

survival curves in Fig. 2-4 further confirm this observation.



(a) PFS.

(b) OS.

Figure 2-4: Kaplan-Meier survival curves broken down by treatment group.

2.3 Methods

2.3.1 Stochastic Model for Tumor Growth

We assume that net tumor growth in our dataset (that is, net of natural tumor growth and drug-induced regression) follows an exponential model. This is one of the most commonly used models in the literature to describe macroscopic tumor growth [34, 35, 36, 37]. We make the rate constant a linear function of the treatment group (as defined in Section 2.2.4), demographics, medical history, and laboratory test features in the dataset (see Table 2.2). This will allow us to identify the key factors driving tumor response. In addition, we incorporate additive Gaussian white noise into the rate constant to account for any randomness intrinsic to the tumor growth process, uncertainty in tumor measurements, or possible errors due to model misspecification.

$$\begin{aligned} \mathrm{d}Y(t) &= \Big(\mu + \boldsymbol{\beta}^\top \boldsymbol{x} + \sigma\,\xi(t)\Big)\,Y(t)\;\mathrm{d}t\,, \qquad\qquad Y(t)\,,\ \sigma > 0\,,\ \boldsymbol{\beta}\,,\ \boldsymbol{x} \in \mathbb{R}^m \\ &= \Big(\mu + \boldsymbol{\beta}^\top \boldsymbol{x}\Big)\,Y(t)\;\mathrm{d}t + \sigma\,Y(t)\;\mathrm{d}W(t) \end{aligned} \tag{2.1}$$

where $Y(t)$ is the SLD measurement[1] for some patient at time t, $\boldsymbol{x}$ is the feature vector for the patient with coefficients $\boldsymbol{\beta}$, m is the number of features, μ is the intercept constant, $(\mu + \boldsymbol{\beta}^\top \boldsymbol{x})$ is the linear function for the net rate of tumor growth, $\sigma\,\xi(t)$ is Gaussian white noise scaled by a positive constant σ, and $W(t)$ is the Wiener process.

The resulting model in Eq. (2.1) is a stochastic differential equation that corresponds to geometric Brownian motion. By applying Itô's formula, we obtain an analytic solution (see Eq. (2.2)). Under the model, the tumor growth process is continuous and can only assume positive values. This agrees well with the physical characteristics of lesions. Since Brownian motion is a Markov process, timepoint

[1] We add a small $\epsilon = 0.1$ mm to cases where SLDs are observed to be 0. We assume that target lesions do not disappear completely under complete response. Instead, they shrink to small sizes that are not easily discovered by the human eye through CT scans or X-ray imaging. We make this assumption so that all SLDs are strictly positive and the data fulfills the $Y(t) > 0$ constraint in the model. It should make no material difference to our analysis, since the ϵ added is substantially smaller than the smallest non-zero SLD observed—1 mm—in the dataset.

transitions—the percentage changes in SLD from one visit to the next—are independent log-normal random variables (see Eq. (2.3)). In other words, tumor sizes in the future are independent of past measurements given the present state. The net rate of tumor growth here, $(\mu + \boldsymbol{\beta}^\top \boldsymbol{x})$, is better known as drift in literature, and σ as volatility.

$$Y(t) = Y(0)\, \exp\Big((\mu + \boldsymbol{\beta}^\top \boldsymbol{x} - \frac{1}{2}\sigma^2)\, t + \sigma\, W(t)\Big) \tag{2.2}$$

$$\ln\Big(\frac{Y(t+h)}{Y(t)}\Big) \sim \mathcal{N}\Big((\mu + \boldsymbol{\beta}^\top \boldsymbol{x} - \frac{1}{2}\sigma^2)\, h\,,\, \sigma^2\, h\Big) \tag{2.3}$$

where $Y(0)$ is the BSLD, h is the time interval between two visits.

As described earlier, patient discontinuation is a complex process that depends on a multitude of factors. At each visit, investigators decide whether or not to discontinue patients based on patient conduct and clinical condition. We propose to model the process as a sequence of Bernoulli trials where patients have a probability of being discontinued at each visit. This probability is conditioned on each individual's target lesion response at the current assessment—that is, whether SLD indicates disease progression, one of the most important factors determining discontinuation—as derived from observed SLD measurements (see Eq. (2.4)). Patients who discontinue move to a state with a null measurement; patients who stay advance to the next visit.

$$D(t) \mid S(t) = s(t) \sim \text{Bernoulli}(P_{s(t)})\,, \quad D(t) \in \{0\,,\, 1\}\,,\ S(t) \in \{\text{PD}\,,\, \text{NPD}\} \tag{2.4}$$

where $D(t)$ is the discontinuation decision at time t, $D(t) = 1$ refers to discontinuation, $D(t) = 0$ refers to continuation, $S(t)$ is the response status of target lesions at time t, and NPD refers to non-PD. We note that $s(t)$ is derived from all past SLDs including the current assessment at time t.

We combine the tumor growth and the patient discontinuation models to obtain the probability density function for our dataset (see Eq. (2.5)). The corresponding likelihood function is shown in Eq. (2.6). We can estimate parameters of both models jointly through maximum likelihood estimation. The set of features for consideration is large (see Table 2.2). Therefore, we employ the stepwise forward selection

algorithm with Akaike information criterion (AIC) as selection criteria to identify a parsimonious set of factors for our model.

$$p(y(t+h) \mid y(t)\,,\, s(t)\,;\, \theta\,, P_{\text{PD}}\,,\, P_{\text{NPD}}) = \begin{cases} P_{s(t)} & \text{for } y(t+h) = \text{NULL} \\ (1 - P_{s(t)}) \cdot f(y(t+h) \mid y(t)\,;\, \theta) & \text{for } y(t+h) > 0 \\ 0 & \text{otherwise} \end{cases} \tag{2.5}$$

$$\mathcal{L}(\theta\,, P_{\text{PD}}\,,\, P_{\text{NPD}} \mid \mathcal{D}) = \prod_{i=1}^{n} \prod_{j=0}^{k^i} p(y^i(t^i_{j+1}) \mid y^i(t^i_j)\,,\, s^i(t^i_j)\,;\, \theta\,, P_{\text{PD}}\,,\, P_{\text{NPD}}) \tag{2.6}$$

where $p(\cdot)$ is the probability density function for our dataset, NULL represents the lack of tumor assessment due to discontinuation, $f(\cdot)$ is the probability density function described in Eq. (2.3), θ are the parameters of the tumor growth model, $\mathcal{L}(\cdot)$ is the likelihood function, $\mathcal{D}$ is the observed data, n is the number of patients in the dataset, k^i is the number of visits for patient i, $y^i(t)$ is the SLD measurement for patient i at time t, t^i_j refers to the time of visit j for patient i, t^i_0 refers to the baseline visit for patient i, $t^i_{k^i+1}$ refers to the cutoff pseudo-visit with null measurement due to discontinuation, and $s^i(t)$ is the response status of patient i at time t. Since the data is cut off at some point for everyone, every patient has a cutoff pseudo-visit. We note that pseudo-visits are not associated with specific times, and the probability of transition to cutoff is independent of time intervals as shown in Eq. (2.5).

The proposed model can be used to simulate tumor growth in patients with different intrinsic characteristics and under different types of therapy. By aggregating outcomes from multiple bootstrap simulations, we can predict the probability of objective response (OR) in patients. This can serve as a baseline predictive model for comparison with the machine-learning models we explore in Section 2.3.2. At the same time, the estimated parameters of the model can provide valuable insights into the drivers behind tumor response that are not readily apparent in the objective response rate. We implement the model in Python.

2.3.2 Machine Learning Models for Objective Response

The prediction of tumor response can be formulated as a supervised binary classification problem, the goal of which is to predict the probability of objective response in individual patients given a set of input features, including the treatment group, demographics, medical history, and baseline laboratory test variables. We consider a total of 45 categorical variables and 5 continuous variables. We convert categorical features to binary variables, and we standardize the continuous variables by their means and standard deviations (see Appendix A.3).

We randomly split the dataset into two disjoint sets, one training set (70%, n=5,463) and one testing set (30%, n=2,342). We use the training set to develop predictive models, and we keep the testing set as an out-of-sample dataset for model validation. In addition to the proposed statistical model, we explore several linear and non-linear machine-learning algorithms commonly used in the literature: penalized logistic regression, decision trees, random forests, and multi-layer perceptrons.

We implement the models in Python using the scikit-learn [38] package and tune their hyper-parameters using five-fold cross-validation with the standard "area under the receiver operating characteristic curve" (AUC) as the metric for model performance. In this context, AUC is the estimated probability that a classifier will rank a responder higher than a non-responder [39]. An AUC of 0.5 corresponds to a random classifier, while 1.0 corresponds to a perfect classifier. Here, we repeat the experiment 100 times for each model to obtain confidence intervals for the expected performance on unseen data. To gain insight into the drivers behind tumor response, we extract and examine the top 20 most important variables of the best-performing predictive model.

Since patients with different biomarker mutations may have different genotypes and phenotypes, and therefore distinct drivers for tumor response under inhibitor therapy, we might expect classifiers trained on treatment-group specific data to outperform the general models trained on the entire dataset. We build and analyze such specialized models by filtering the dataset by treatment group prior to training and

testing. Subsequently, we compute an AUC for the entire dataset by aggregating predictions from all six specialized models (one for each treatment group). As a comparison, we also break down the performance of the general models by treatment group.

2.3.3 Statistical Models for Survival

The modeling of time-to-event data such as PFS and OS is known as survival analysis. It is often performed in medical settings to identify groups at risk of adverse clinical events. At its core, survival analysis is a regression problem involving the estimation of the time (or risk) of occurrence of some event of interest, given a set of dependent variables (see Appendix A.3). One of the main challenges in modeling such longitudinal data is the presence of censoring. Censoring occurs when the event of interest fails to happen within the finite span of the clinical trial, and the patient is lost to follow-up for reasons unrelated to the event. In general, standard regression techniques are not able to accommodate this data structure.

In this chapter, we examine two standard methods used in survival modeling: the Cox proportional hazards model [40] and the accelerated failure time model with log-normal distribution. As with the statistical model for tumor growth, we perform stepwise forward selection with AIC as criteria to identify a parsimonious set of features for each of the two models. In addition, we explore two nonlinear and non-parametric survival models: the random survival forest model [41] and the neural network survival model [42]. We implement the Cox, accelerated failure time, and random survival forest models in R using packages survival [43] and randomForestSRC [44], while we implement the neural network survival model in Python using PyTorch. For the PFS models, we consider features in the same way as in the models for tumor response. In view of the confounding effects of crossovers and post-trial therapies, we replace the treatment group feature with biomarker positivity in PDL1, EGFR, or ALK for the OS models (see Appendix A.3).

We adopt the training and validation methodology described in Section 2.3.2. Instead of the AUC, we use the concordance index (C-index) as the metric for model

performance. The C-index is commonly used to evaluate predictive power in survival analysis. It is basically a measure of the concordance between orderings of observed survival times and predicted times or risks [45]. Like AUCs, C-indices can range between 0.5 (random predictions) and 1.0 (perfect model).

In addition to their discriminative power, we assess model calibration by comparing the actual and the predicted survival probabilities at different times: 6, 12 and 24 months for PFS; and 12, 24 and 36 months for OS. For each time cutoff, we divide the test set into quintiles based on the predicted risk scores. We then compute the average predicted score and the true survival probability observed in each of the quintiles. Finally, we obtain calibration plots by plotting the observed probabilities against the predicted probabilities. In the ideal case, the points will lie as close as possible to the diagonal line, which represents perfect calibration. Lastly, we stratify patients in the test set into different risk groups based on their predicted risk scores, and we examine the differences in their Kaplan-Meier survival curves using the log-rank test.

2.4 Results

2.4.1 Tumor Response

We summarize the AUC results for the test set in Table 2.7. In general, we find that the machine-learning models substantially outperform the baseline statistical model, with improvements up to 0.07 AUC. In particular, logistic regression achieves the best performance out of all models, with 0.79 AUC (95% C.I. 0.77–0.81). We do not observe any appreciable difference in predictive power between the general models and the specialized models. We note that the specialized random forest models demonstrate marginally better performance than the general logistic regression model. Nevertheless, we favor the latter due to its ease of interpretability.

The breakdown in performance by treatment group reveals a wide spread in predictive powers. For example, the general logistic regression model is capable of AUCs as high as 0.8 for the untested treatment group under inhibitor therapy, but manages a

performance that is only slightly better than random (0.53 AUC) for the group tested negatively for biomarkers but under inhibitor therapy. This phenomenon persists in the specialized models as well.

We extract the top 20 largest coefficients—that is, most important variables—of the general statistical model and the best-performing general logistic regression model. We compute the average estimates over all experiments and summarize their standard deviations in Tables 2.8 and 2.9. The parameters are generally estimated with good precision.

The probability of discontinuation estimated in the statistical model indicate an 81% chance of discontinuation given disease progression in target lesions at the current visit. The remaining 19% accounts for the possibilities of follow-up visits and treatment beyond progression at the investigators' discretion. Conditional on a lack of disease progression in target lesions, there is a 16% chance of data cutoff due to progression in non-target lesions, appearance of new lesions or other adverse events. Otherwise, the patient has an 84% chance of advancing to the next visit.

In Tables 2.8 and 2.9, we rank the coefficients according to the extent of their effects on tumor growth and the probability of OR. In both the statistical model and the logistic regression model, we find biomarker status to be among the strongest drivers of regression and response in patients treated with immunotherapies and targeted therapies. It is also interesting to note that EGFR TKI seems to have weaker effects—that is, smaller coefficients—as compared with ALK TKI and PDL1 ICI. We discuss possible reasons for this in Section 2.5.

Patients who have not undergone prior chemotherapy, nonsmokers, and women are more likely to respond to therapy. On the other hand, patients who have squamous cell carcinoma (SCC) histology or who suffer from gastrointestinal comorbidities are at greater risk for progression. Not all factors overlap in both models. Here, we focus on the logistic regression model, which has demonstrated stronger predictive powers. Under this model, Asians have a greater likelihood of OR than non-Asians, and patients that have normal functional status—ECOG performance status 0—have a greater likelihood of OR than physically impaired patients. Other important risk

factors include low hemoglobin count, elevated alanine aminotransferase and alkaline phosphate levels, trial enrollment in Western Europe, and liver metastasis.

Table 2.7: Performance of predictive models for tumor response broken down by treatment group. General models are trained on the entire dataset. Specialized models are trained on treatment-group specific data.

	Test Set Average AUC (95% C.I.)				
	Statistical Model	**Logistic Regression**	**Decision Tree**	**Random Forest**	**Multi-layer Perceptron**
General Model					
All	0.716 (0.695, 0.740)	0.787 (0.770, 0.805)	0.736 (0.718, 0.750)	0.774 (0.759, 0.791)	0.779 (0.764, 0.795)
Chemotherapy	0.628 (0.594, 0.670)	0.707 (0.675, 0.740)	0.648 (0.611, 0.685)	0.706 (0.678, 0.742)	0.694 (0.661, 0.729)
PDL1-positive and under PDL1 ICI	0.522 (0.459, 0.589)	0.561 (0.507, 0.620)	0.536 (0.484, 0.597)	0.556 (0.494, 0.613)	0.556 (0.503, 0.607)
EGFR-positive and under EGFR TKI	0.583 (0.517, 0.645)	0.719 (0.669, 0.764)	0.652 (0.568, 0.715)	0.709 (0.663, 0.753)	0.690 (0.634, 0.746)
ALK-positive and under ALK TKI	0.550 (0.482, 0.614)	0.625 (0.575, 0.685)	0.551 (0.493, 0.616)	0.620 (0.571, 0.678)	0.622 (0.561, 0.685)
Negative biomarker but under inhibitor therapy	0.504 (0.436, 0.581)	0.530 (0.458, 0.595)	0.547 (0.486, 0.616)	0.507 (0.437, 0.583)	0.547 (0.477, 0.632)
Not tested but under inhibitor therapy	0.739 (0.695, 0.784)	0.813 (0.777, 0.853)	0.748 (0.709, 0.793)	0.811 (0.771, 0.851)	0.803 (0.766, 0.838)
Specialized Models					
All[1]	0.724 (0.707, 0.738)	0.775 (0.749, 0.789)	0.754 (0.718, 0.772)	0.790 (0.784, 0.800)	0.736 (0.700, 0.764)
Chemotherapy	0.627 (0.581, 0.661)	0.706 (0.676, 0.736)	0.651 (0.627, 0.681)	0.704 (0.677, 0.734)	0.658 (0.622, 0.692)
PDL1-positive and under PDL1 ICI	0.516 (0.461, 0.579)	0.528 (0.467, 0.581)	0.515 (0.453, 0.586)	0.555 (0.495, 0.609)	0.547 (0.479, 0.613)
EGFR-positive and under EGFR TKI	0.515 (0.419, 0.591)	0.690 (0.643, 0.732)	0.664 (0.623, 0.706)	0.717 (0.684, 0.753)	0.657 (0.577, 0.718)
ALK-positive and under ALK TKI	0.478 (0.409, 0.541)	0.605 (0.544, 0.672)	0.548 (0.479, 0.607)	0.609 (0.557, 0.657)	0.566 (0.509, 0.632)
Negative biomarker but under inhibitor therapy	0.563 (0.485, 0.629)	0.561 (0.491, 0.632)	0.530 (0.432, 0.605)	0.580 (0.511, 0.640)	0.551 (0.470, 0.621)
Not tested but under inhibitor therapy	0.721 (0.660, 0.776)	0.797 (0.766, 0.838)	0.782 (0.750, 0.811)	0.807 (0.778, 0.839)	0.779 (0.745, 0.817)

[1] C-index derived from predictions aggregated from specialized models.

2.4.2 Survival

We summarize the test set C-index results in Tables 2.10 and 2.11. We find the performance to be similar across linear and nonlinear models for both PFS and OS. This suggests that the prediction problem does not necessarily benefit from models with greater complexity. We focus on the Cox model here due to its ease of implementation, application, and interpretability. The model achieves C-indices of 0.67 (95% C.I. 0.66–0.69) and 0.73 (95% C.I. 0.72–0.74) on out-of-sample data for PFS and OS, respectively. Our experiments indicate that specialized PFS models perform more poorly than their general counterparts. As in Section 2.4.1, we observe a spread in discriminative powers among the treatment groups, albeit less pronounced, with the PDL1-positive group under PDL1 ICI treatment and the negative biomarker group under inhibitor therapy having the lowest C-indices.

We show calibration plots of the Cox models at multiple time points in Figs. 2-5 and 2-6. The calibration curves lie close to the ideal diagonal, indicating that the models are well-calibrated. They do not systematically overestimate or underestimate survival rates in any of the quintiles. We only observe larger confidence intervals at longer survival times, which is expected given that the number of patients at risk decreases over time. We further stratify patients into two risk groups using a median split of the risk scores, and plot their survival curves in Fig. 2-7. The difference in survival between the high and low risk groups is significant (log-rank test $p < 0.0001$) for both PFS and OS. In general, the low-risk patients have longer median survival times than the high-risk patients.

In Tables 2.12 and 2.13, we extract the top 20 coefficients in the Cox models to identify specific factors that predict PFS and OS. The parameters are estimated with good precision. We find biomarker positivity to be associated with better PFS for patients treated with inhibitor therapies. According to the PFS Cox model, a performance status of 2 or higher, metastasis in liver, elevated white blood cells and alanine aminotransferase levels, histological subtype SCC, low hemoglobin count, and gastrointestinal and blood-related comorbidities are related to increased risk.

Table 2.8: Probability of discontinuation and top 20 coefficients of the general statistical model. Factors with negative coefficients drive tumor regression, while factors with positive coefficients have the opposite effect.

Type	**Variable**	**Average Estimate**	**Standard Deviation**
Discontinuation Model			
Probability of discontinuation given PD in target lesions at current visit	P_{PD}	0.809	0.006
Probability of discontinuation given NPD in target lesions at current visit	P_{NPD}	0.159	0.001
Exponential Model		(10-3)	(10^{-3})
Treatment group	ALK-positive and under ALK TKI	-1.497	0.109
Intercept	μ	1.337	0.157
Treatment group	PDL1-positive and under PDL1 ICI	-1.116	0.125
Medical history	Prior chemotherapy - No	-0.823	0.101
Laboratory measurements	Bilirubin - Low	-0.683	0.443
Medical history	Histology - SCC	0.544	0.105
Medical history	Number of baseline target lesions - 5 or more	-0.431	0.263
Demographics	Race group - Others	-0.418	0.373
Laboratory measurements	Alkaline phosphate - Low	-0.380	0.498
Medical history	Smoking status - Never	-0.371	0.119
Laboratory measurements	Creatine - High	-0.355	0.244
Comorbidities	Endocrine disorders - Yes	-0.278	0.270
Demographics	Sex - Female	-0.273	0.119
Metastasis	Others - No	0.239	0.313
Comorbidities	Social circumstances - Yes	-0.239	0.341
Treatment group	EGFR-positive and under EGFR TKI	-0.220	0.190
Comorbidities	Gastrointestinal disorders - Yes	0.194	0.130
Metastasis	Number of metastasis sites	-0.169	0.064
Laboratory measurements	White blood cells count - High	-0.151	0.159
Medical history	Histology - Others	0.139	0.198
Comorbidities	Respiratory, thoracic and mediastinal disorders - Yes	-0.113	0.121

Table 2.9: Top 20 coefficients of the general logistic regression model. Factors with positive coefficients improve the odds of response, while factors with negative coefficients have the opposite effect.

Type	**Variable**	**Average Coefficient**	**Standard Deviation**
Intercept	Intercept	-1.692	0.347
Treatment group	ALK-positive and under ALK TKI	1.649	0.316
Treatment group	PDL1-positive and under PDL1 ICI	1.118	0.275
Medical history	Prior chemotherapy - No	0.837	0.111
Treatment group	EGFR-positive and under EGFR TKI	0.663	0.157
Medical history	Histology - SCC	-0.436	0.069
Laboratory measurements	Hemoglobin - Low	-0.204	0.050
Medical history	Performance status - 2 or higher	-0.199	0.076
Medical history	Smoking status - Never	0.197	0.051
Metastasis	Liver - Yes	-0.175	0.052
Demographics	Sex - Female	0.168	0.060
Laboratory measurements	Alanine aminotransferase - High	-0.145	0.068
Demographics	Region - WEUR	-0.142	0.076
Medical history	Performance status - 0	0.140	0.045
Treatment group	Not tested but under inhibitor therapy	-0.130	0.094
Laboratory measurements	White blood cells count - High	-0.124	0.052
Comorbidities	Gastrointestinal disorders - Yes	-0.121	0.044
Laboratory measurements	Alkaline phosphate - High	-0.117	0.040
Demographics	Race group - Asian	0.104	0.083
Demographics	Region - Others	-0.092	0.080
Metastasis	Brain - Yes	-0.078	0.055

Conversely, women, nonsmokers, patients not previously treated with chemotherapy and those with low aspartate aminotransferase levels and normal functional status are more likely to have a positive prognosis.

For OS, we find that the presence of a proven biomarker mutation in PDL1, EGFR, or ALK leads to improved survival. Other factors that have positive effects on OS include the lack of prior chemotherapy exposure, Asian ethnicity, a performance status of 0, no history of smoking, and female sex. Adverse risk factors include high white blood cell count, histological subtype SCC, low hemoglobin level, liver metastasis, and BSLD size. We note that many of the features here also appear in the PFS model, and have similar effects on survival.

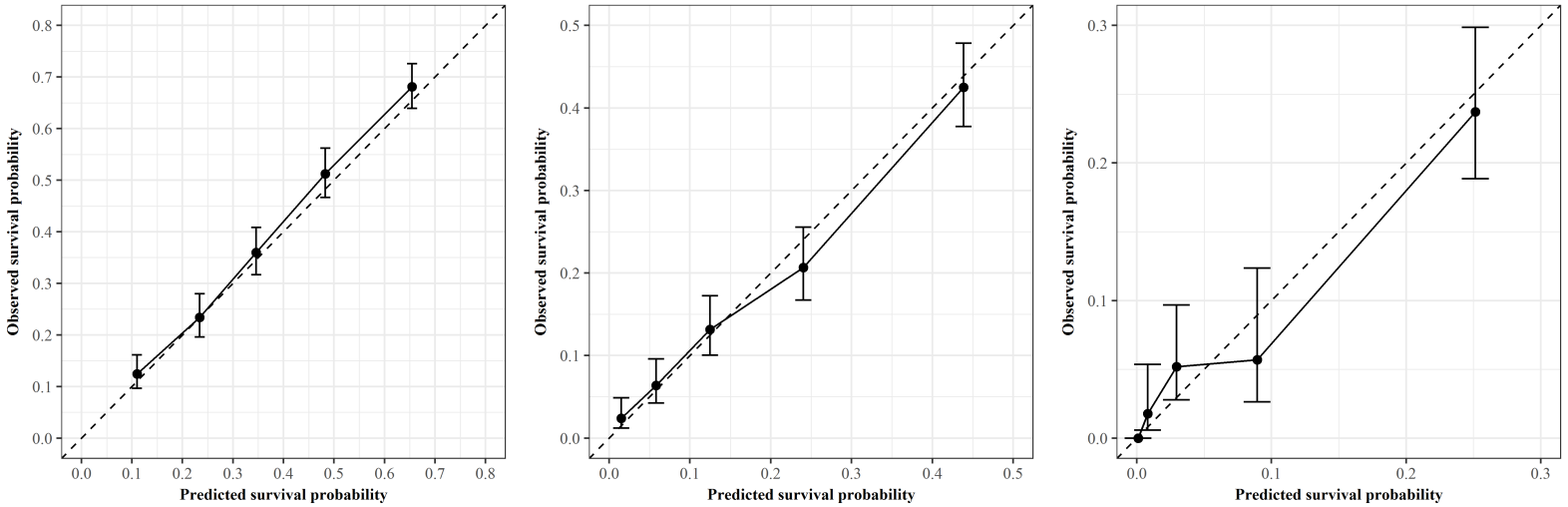
Table 2.10: Performance of predictive models for PFS broken down by treatment group.

	Test Set Average C-index (95% C.I.)			
	Cox Proportional Hazards	**Accelerated Failure Time**	**Random Survival Forest**	**Multi-layer Perceptron**
General Model				
All	0.673 (0.664, 0.685)	0.670 (0.660, 0.682)	0.672 (0.663, 0.684)	0.678 (0.668, 0.689)
Chemotherapy	0.660 (0.646, 0.675)	0.660 (0.645, 0.674)	0.650 (0.634, 0.663)	0.660 (0.645, 0.673)
PDL1-positive and under PDL1 ICI	0.602 (0.568, 0.639)	0.601 (0.561, 0.636)	0.592 (0.554, 0.635)	0.607 (0.568, 0.647)
EGFR-positive and under EGFR TKI	0.693 (0.667, 0.722)	0.699 (0.672, 0.728)	0.707 (0.678, 0.737)	0.703 (0.676, 0.734)
ALK-positive and under ALK TKI	0.654 (0.619, 0.690)	0.648 (0.614, 0.684)	0.646 (0.610, 0.685)	0.655 (0.615, 0.692)
Negative biomarker but under inhibitor therapy	0.603 (0.577, 0.630)	0.600 (0.573, 0.629)	0.606 (0.576, 0.636)	0.617 (0.591, 0.642)
Not tested but under inhibitor therapy	0.642 (0.626, 0.658)	0.638 (0.620, 0.654)	0.639 (0.619, 0.658)	0.649 (0.632, 0.668)
Specialized Models				
All[1]	0.629 (0.620, 0.639)	0.660 (0.651, 0.668)	0.618 (0.599, 0.646)	0.628 (0.614, 0.641)
Chemotherapy	0.651 (0.636, 0.667)	0.651 (0.635, 0.669)	0.649 (0.632, 0.666)	0.640 (0.626, 0.658)
PDL1-positive and under PDL1 ICI	0.561 (0.516, 0.603)	0.550 (0.498, 0.590)	0.572 (0.532, 0.607)	0.558 (0.517, 0.598)
EGFR-positive and under EGFR TKI	0.677 (0.646, 0.708)	0.670 (0.636, 0.705)	0.696 (0.667, 0.721)	0.661 (0.623, 0.693)
ALK-positive and under ALK TKI	0.631 (0.591, 0.664)	0.628 (0.587, 0.664)	0.637 (0.599, 0.675)	0.615 (0.580, 0.645)
Negative biomarker but under inhibitor therapy	0.597 (0.571, 0.621)	0.588 (0.560, 0.618)	0.603 (0.577, 0.628)	0.572 (0.546, 0.604)
Not tested but under inhibitor therapy	0.636 (0.616, 0.655)	0.633 (0.610, 0.653)	0.638 (0.616, 0.659)	0.625 (0.599, 0.647)

[1] C-index derived from predictions aggregated from specialized models.

Table 2.11: Performance of predictive models for OS.

	Test Set Average C-index (95% C.I.)			
	Cox Proportional Hazards	**Accelerated Failure Time**	**Random Survival Forest**	**Multi-layer Perceptron**
General Model				
All	0.729 (0.721, 0.739)	0.726 (0.717, 0.736)	0.725 (0.715, 0.734)	0.729 (0.720, 0.739)

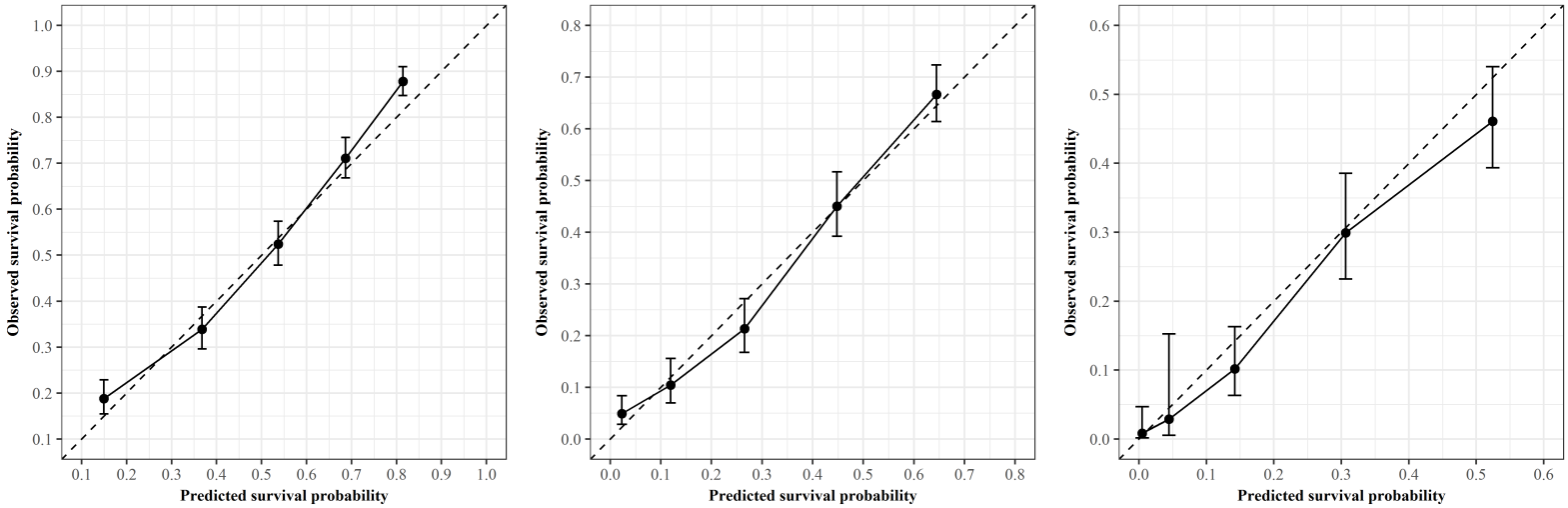


(a) At 6 months.

(b) At 12 months.

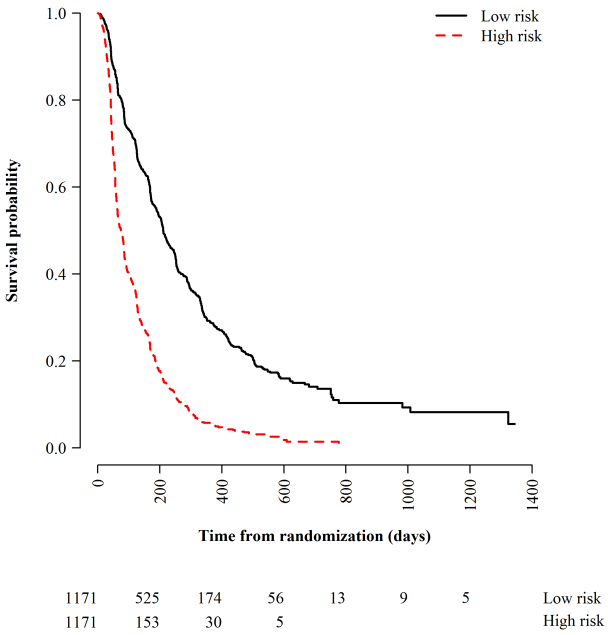
(c) At 24 months.

Figure 2-5: Calibration plots of the PFS Cox proportional hazards model for a random experiment iteration. Dashed line represents the ideal model. Vertical bars show the 95% confidence intervals.

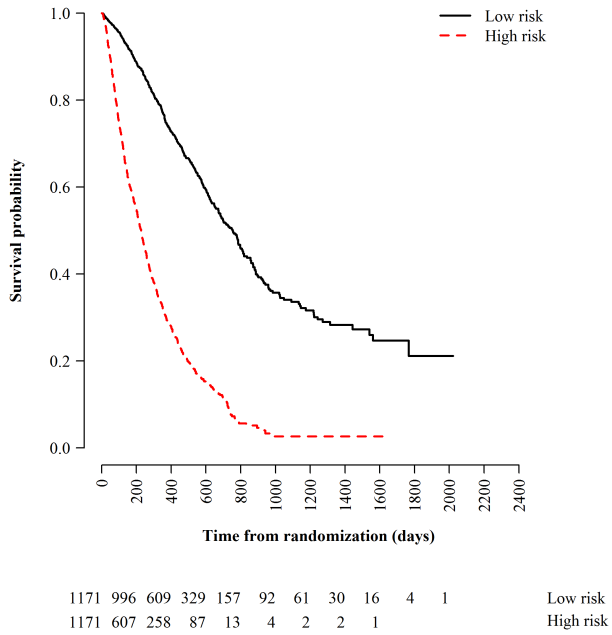


(a) At 12 months. (b) At 24 months. (c) At 36 months.

Figure 2-6: Calibration plots of the OS Cox proportional hazards model for a random experiment iteration. Dashed line represents the ideal model. Vertical bars show the 95% confidence intervals.



(a) Risk stratification by the PFS Cox proportional hazards model. Hazard ratio 2.58 (95% C.I. 2.35–2.85). Log-rank test $p < 0.0001$.



(b) Risk stratification by the OS Cox proportional hazards model. Hazard ratio 3.94 (95% C.I. 3.52–4.42). Log-rank test $p < 0.0001$.

Figure 2-7: Risk stratification by survival models for a random experiment iteration.

Table 2.12: Top 20 coefficients of the PFS Cox proportional hazards model. Factors with negative coefficients improve survival, while factors with positive coefficients have the opposite effect.

Type	Variable	Average Coefficient	Standard Deviation
Treatment group	ALK-positive and under ALK TKI	-0.928	0.040
Treatment group	PDL1-positive and under PDL1 ICI	-0.787	0.047
Medical history	Prior chemotherapy - No	-0.487	0.028
Treatment group	EGFR-positive and under EGFR TKI	-0.393	0.038
Medical history	Performance status - 2 or higher	0.370	0.043
Metastasis	Liver - Yes	0.242	0.028
Laboratory measurements	White blood cells count - High	0.222	0.024
Laboratory measurements	Bilirubin - High	0.182	0.127
Laboratory measurements	Alkaline phosphate - Low	0.173	0.153
Metastasis	Others - No	0.166	0.066
Laboratory measurements	Aspartate aminotransferase - Low	-0.145	0.169
Laboratory measurements	Alanine aminotransferase - High	0.134	0.063
Demographics	Race group - Others	-0.133	0.101
Demographics	Sex - Female	-0.131	0.025
Medical history	Histology - SCC	0.128	0.035
Laboratory measurements	Hemoglobin - Low	0.125	0.022
Comorbidities	Blood and lymphatic system disorders - Yes	0.122	0.048
Medical history	Performance status - 0	-0.115	0.018
Comorbidities	Gastrointestinal disorders - Yes	0.113	0.023
Medical history	Smoking status - Never	-0.113	0.035

2.5 Discussion

We have developed predictive models for OR, PFS, and OS in advanced-NSCLC patients receiving chemotherapy, targeted therapy and immunotherapy. The models are built on a large dataset composed of 7,805 patients from 17 clinical trials, which have recorded a wide range of tumor-related and patient-related factors routinely collected before clinical testing. They demonstrate promising predictive abilities, achieving out-of-sample performances of 0.79 AUC, 0.67 C-index, and 0.73 C-index for OR, PFS, and OS, respectively. The calibration plots suggest good agreement between the actual and predicted survival probabilities.

By examining the top coefficients in the linear predictive models, we identify the specific factors that have the greatest predictive values for tumor response and survival. Since the endpoints for OR, PFS, and OS are related to one another[2], many of the important features found are common across the models. From Tables 2.9 and 2.12, we observe treatment group to be among the strongest predictors of response

[2]PFS is dependent on both response and mortality data, from which OR and OS are derived, respectively.

Table 2.13: Top 20 coefficients of the OS Cox proportional hazards model. Factors with negative coefficients improve survival, while factors with positive coefficients have the opposite effect.

Type	Variable	Average Coefficient	Standard Deviation
Medical history	Performance status - 2 or higher	0.601	0.049
Medical history	Prior chemotherapy - No	-0.512	0.034
Biomarker positivity	PDL1, EGFR or ALK - Yes	-0.481	0.028
Demographics	Race group - Asian	-0.371	0.088
Medical history	Performance status - 0	-0.369	0.021
Laboratory measurements	White blood cells count - High	0.335	0.029
Laboratory measurements	Bilirubin - High	0.296	0.129
Medical history	Histology - SCC	0.263	0.034
Laboratory measurements	Alkaline phosphate - High	0.253	0.026
Laboratory measurements	Hemoglobin - Low	0.237	0.024
Metastasis	Liver - Yes	0.218	0.028
Medical history	Smoking status - Never	-0.200	0.026
Comorbidities	Eye disorders - Yes	-0.178	0.059
Demographics	Sex - Female	-0.175	0.025
Medical history	BSLD	0.173	0.014
Medical history	Histology - Others	0.150	0.044
Comorbidities	Nervous system disorders - Yes	-0.134	0.033
Comorbidities	Renal and urinary disorders - Yes	-0.133	0.084
Demographics	Weight	-0.125	0.013
Comorbidities	Skin and subcutaneous tissue disorders - Yes	-0.123	0.061

and survival endpoints. In particular, biomarker mutation status has a substantial effect on the efficacy of inhibitor therapies. Among patients treated with ICI and TKI, those that tested positively for biomarkers have a higher probability of OR and a more favorable PFS than those who either tested negatively or are untested. Patients in the former treatment group also tend to have a more optimistic prognosis than patients treated with chemotherapy in general. Our findings are consistent with that observed in the ORR (see Table 2.5) and the survival data (see Table 2.6 and Fig. 2-4). For OS, we find that the presence of an actionable mutation in PDL1, EGFR, or ALK is associated with improved survival. This likely reflects survival benefits from advances in molecularly targeted therapy and immunotherapy that are based on these three driver mutations.

As reported in many studies, we find that nonsmokers, women, and those with good performance status have a higher chance of responding to treatment [46, 47, 48, 49]. In addition to these well-established prognostic factors, we find prior chemotherapy to be a strong predictor: patients who have undergone prior chemotherapy treatments have poorer survival than chemotherapy-naive patients. The failure of

chemotherapy suggests that the tumor cells are to some extent drug resistant and more difficult to treat. Consequently, patients that do not go into remission after chemotherapy are less likely to respond to subsequent treatments. Our models did not identify either age or stage of cancer as significant features, despite their well-established status as prognostic factors. We believe that this is because the dataset is composed principally of advanced-stage patients who are relatively close in age to each other.

We find that patients with abnormal baseline laboratory measurements tend to have a poorer prognosis. Irregular blood test results are often indicative of underlying physiological disorders that may interfere with cancer treatment delivery and thus adversely affect survival [50]. For example, elevated alanine aminotransferase and bilirubin levels are related to possible liver damage and hepatic dysfunction, while a low hemoglobin count is associated with anemia and cachexia, and a high white blood cell count is linked to systemic inflammation and subclinical infection. This underscores the importance of comorbidity management to improve survival, and also calls for the collection of more data on such conditions so that the experience of these patients can be adequately analyzed through appropriate stratification in clinical studies. Consistent with other studies, we identify liver metastasis [51] and SCC histology as important risk factors. We note that driver mutations EGFR and ALK are rarely found in SCC [52]. Thus, this group of patients generally have worse clinical outcomes than non-squamous patients because they have fewer effective treatment options.

In Table 2.9, we find that the coefficient for patients who are PDL1-positive and under PDL1 ICI treatment is almost double that of the analogous coefficient for EGFR-positive patients under EGFR TKI. This seems unusual, given that patients in the former treatment group have a lower ORR than those in the latter (see Table 2.5). A closer look at the testing set AUCs in Table 2.7 reveals that the logistic regression model does not perform as well on the PDL1-positive treatment group (0.56 AUC 95% C.I. 0.51–0.62) as on the EGFR-positive subgroup (0.72 AUC 95% C.I. 0.67–0.76). This suggests that the prognostic factors described above are not as

relevant for PDL1-positive patients that are under PDL1 ICI. Since we observe the same trend in specialized and nonlinear models, we know that the poor performance is not due to the linearity of the applied model, but rather due to the lack of strong predictors for response within the PDL1-positive under PDL1 ICI subgroup. Therefore, in the absence of other useful factors, the algorithm attributes the higher than average ORR in the subgroup mainly to the effectiveness of ICI on PDL1-positive patients, and assigns greater weight to the corresponding treatment group indicator so that it dominates other factors in the dataset. This also in part explains why the ALK-positive and under ALK TKI treatment group has such a large coefficient. In contrast, because much of the response in the EGFR-positive subgroup can be explained away by other factors, the corresponding treatment group coefficient has a smaller magnitude. The same holds for the PFS Cox model.

The importance of PDL1 positivity for response to PDL1 ICI is well established. However, our results suggest that there other factors beyond PDL1 expression at play since not all PDL1-positive patients responded to treatment. These additional variables seem to be absent from the current feature set. Having ruled out demographic, clinical, and pathological factors that are already present in the dataset, we suspect that PDL1-positive patients may possess additional biomarkers, such as germline mutations, that affect treatment efficacy and predispose patients to specific responses. In fact, there is emerging evidence that variables such as tumor-mutational burden are predictive of response to immunotherapy [53]. Unfortunately, while genomic profiling is now almost routine in clinical trials, such data is typically not examined holistically and rarely submitted to the FDA.

To the best of our knowledge, our work is one of the largest studies of NSCLC to consider biomarker mutation and inhibitor therapy as candidate predictive variables. With the emergence of new treatment pathways that significantly improve survival in patients with relevant biomarkers, the importance of mutation status as a predictive factor cannot be understated [54]. Putila et al. [47] developed a prognostic model for OS based on almost two decades of data (1998-2006) from the Surveillance, Epidemiology, and End Results Program (SEER) database. However, while the sample size

is impressive (over 230,000 patients), the dataset does not capture recent advances in cancer treatment, such as targeted therapies. Alexander et al. [46] proposed the Lung Cancer Prognostic Index (LCPI) to predict OS. This model includes actionable mutations in EGFR, ALK, and KRAS as features, but lacks the PDL1 biomarker. Its derivation cohort is also much smaller (roughly 700 patients) than the training set used here (around 5,400 patients). A direct comparison of predictive performances to the SEER and LCPI models is limited by heterogeneity in data and features. Both models focus on prognosis at the time of diagnosis. In contrast, our model focuses on prediction prior to therapy for OR and PFS, and anytime after diagnosis for OS. In addition, the SEER model requires TNM cancer staging information, which is not available in our dataset. We also do not have information on the weight loss at diagnosis used in the LCPI model. In general, we find that a vast majority of prognostic studies focus their attention on OS [55]. Here, we additionally develop similar models for two other clinical endpoints of interest to patients and physicians, OR and PFS.

This study has several limitations. First, the dataset is based on clinical trial data. Clinical trials have strict inclusion and exclusion eligibility criteria, such as a minimum performance status or a specific prior chemotherapy exposure. As a result, the patients enrolled may not be representative of the heterogeneous, real-world patient population [56]. In this study, we try to increase the heterogeneity of our patient cohort by pooling patients from multiple trials. However, advanced-stage patients still make up a large part of the dataset. Second, our models are not tested by populations independent of our study sample. It would be desirable to validate our models with patients outside the clinical trial setting and in the general population. Once validated, the models could be translated into clinical use as a web application, like Adjuvant! Online, one that is easily accessible to patients and physicians.

In this chapter, we aggregate data from 17 clinical trials to estimate OR, PFS, and OS models applicable to patients under different treatment modalities, including chemotherapy, targeted therapy, and immunotherapy. The models include established and novel predictive factors in lung cancer. We offer an interpretation of the effects of each variable and find them to be largely consistent with other NSCLC prognostic

tools in the literature. The models have broad clinical utility in developing individualized treatment plans and augmenting clinical trial design. Our predictions are able to complement the standard protocols used by physicians to guide medical decisions by better informing their patients about their likelihood of response and predicted survival under different therapies. The models illuminate the drivers behind response and survival, which may be potentially useful for patient selection in clinical trials. Survival risk scores can also be used in randomized trials to stratify patients into groups with homogeneous risk for analysis. The predictive results are promising for chemotherapy and targeted therapies, but much less so for immunotherapy. We hypothesize that the lackluster performance in the PDL1-positive subgroup under PDL1 ICI treatment is due to its relatively small sample size and the lack of relevant genomic predictors. It is clear that PDL1 positivity alone does not tell the whole story for immunotherapy.

To advance beyond the results achieved in this chapter, the model must include genomics, immunogenomics, metabolomics, and other composite and complex multiomic signatures that can reflect the state of the tumor, its microenvironment, and the microbiota in general. For example, radiomic features from deep learning models have shown immense potential in NSCLC prognostication [57, 58, 59, 60, 61, 62, 63]. Such variables will inevitably become more important as we gradually reach the limits of biomedical reductionism—that is, the target-based treatment paradigm—and shift toward systems biology approaches for drug discovery [64]. This will allow us to develop better predictive models to truly tailor treatments and clinical trial enrollment strategies.

Chapter 3

Predictive Models for Drug Development Programs

Drug development is an extremely costly process, and the accurate evaluation of a candidate drug's likelihood of approval is critical to the efficient allocation of capital. However, developers typically use general estimates of regulatory approval rates in managing their portfolio of investigational drugs. In this chapter, we propose the use of a wide range of drug and clinical-trial features, and apply machine-learning techniques to predict whether a drug candidate will graduate from phase 2 to approval and from phase 3 to approval. We use two large pharmaceutical pipeline databases to train our models, and apply statistical imputation methods to deal with missingness in the data. We achieve promising levels of predictive power, and find that the most important features for predicting success are trial outcomes, trial status, trial accrual rates, durations, prior approval, and sponsor track records. Our models may be used to evaluate the risks of different investigational drugs at different clinical stages more accurately. Such predictive analytics can be used to make more informed data-driven decisions in risk assessment and portfolio management. This can increase the efficiency of drug development by allowing various stakeholders, including pharmaceutical companies, biotech entrepreneurs, investors, and regulators to better assess the risks and drivers of drug approvals and failures.

3.1 Introduction

While many recent medical breakthroughs such as immuno-therapies, gene therapies, and gene-editing techniques offer new hope for patients, they have also made biomedical innovation riskier, and more complex and expensive. These breakthroughs generate novel therapies for investigation, each of which requires many years of translational research and clinical testing, costing hundreds of millions to billions of dollars, and yet often face a high likelihood of failure [23]. In fact, drug development productivity—the ratio of the number of new drugs approved to R&D spending each year—has declined steadily over the past 50 years despite scientific and technical progress. This phenomenon, which Scannell et al. [1] termed "Eroom's Law," as the reverse of Moore's Law, suggests that the cost of developing new drugs has doubled approximately every nine years since the 1950s. In the face of multiple uncertainties, the need to evaluate drug candidates better and allocate capital to high-potential opportunities more efficiently has only intensified.

To address these needs, in this chapter we apply machine-learning techniques to predict the outcomes of randomized clinical trials. Machine learning is an interdisciplinary field focused on tackling pattern recognition problems and building predictive models to make data-driven decisions, which is well-suited for this context. Successful applications of these techniques have already revolutionized a number of industries (e.g., advertising, marketing, finance and insurance, oil and gas exploration) and are poised for even greater impact via autonomous vehicles, facial-recognition authentication, and general-purpose robotics.

Drug developers have already applied machine-learning tools to the discovery process via high-throughput screening of vast libraries of chemical and biological compounds to identify drug targets. However, in managing their portfolios of investigational drugs, biopharma companies typically use unconditional estimates of regulatory approval rates based on historically observed relative frequencies. Machine learning techniques yield conditional estimates of success, conditioned on a host of predictive factors known to affect the likelihood of approval, including drug compound charac-

teristics, clinical trial design, previous trial outcomes, and the sponsor track record. We show that these features contain useful signals about drug development outcomes that will allow us to forecast the outcome of pipeline developments more accurately.

Our methodology and results have several implications for stakeholders in the biomedical ecosystem. More accurate forecasts of the likelihood of success of clinical trials will reduce the uncertainty surrounding drug development, which will increase the amount of capital that investors and drug developers are willing to allocate to this endeavor. By extension, this would lower the cost of capital and increase the efficiency of the allocation process. Specifically, we predict the probability of success of drug candidates in two scenarios: (1) advancing from phase 2 to regulatory approval and (2) from phase 3 to regulatory approval (see Fig. 3-1). Investors and drug developers may use such predictions to evaluate the risks of different investigational drugs at different clinical stages, providing them with much-needed transparency. Greater risk transparency is one source of improved financial efficiency because it facilitates more accurate matching of investor risk preferences with the risks of biomedical investment opportunities.

Machine-learning models can offer guidance to scientists, clinicians, and biopharma professionals as to which factors are most important in determining clinical-trial success, suggesting ways to improve the drug development process and decelerate or reverse Eroom's Law. Policymakers and regulators would also benefit from machine-learning predictions, particularly for drug-indication pairs that are predicted to fail with high likelihood—these cases highlight the most difficult challenges in biomedicine and underscore the need for greater government and philanthropic support.

To the best of our knowledge, our study is the largest of its kind. We construct two datasets, one for each scenario, from two proprietary pharmaceutical pipeline databases, Pharmaprojects and Trialtrove provided by Informa® [65]. The phase-2-to-approval dataset includes more than 4,000 unique drugs for 288 indications and over 14,500 phase 2 trials, and the phase-3-to-approval dataset contains more than 1,400 unique drugs for 253 indications and over 4,500 phase 3 trials. These data cover over

15 indication groups. In contrast, most published research on drug approval prediction have very small sample sizes, are concentrated on specific therapeutic areas, and involve only one or a small number of predictive factors: Malik et al. [66] examined the trial objective responses of 88 anticancer agents in phase 1; Goffin et al. [67] studied the tumor response rates of 58 cytotoxic agents in 100 phase 1 trials and 46 agents in 499 phase 2 trials; El-Maraghi et al. [68] looked at the objective responses of 19 phase 2 anticancer drugs in 89 single agent trials; Jardim et al. [69] examined the response rates of 80 phase 3 oncology drugs to identify factors associated with failures; and DiMasi et al. [70] analyzed 62 cancer drugs and proposed an approved new drug index (ANDI) algorithm with four factors to predict approval for lead indications in oncology after phase 2 testing (see Appendix B.4 for a comparison of our analysis to theirs).

Another key difference in our approach is that we deal with missing data using statistical imputation methods. We explore four common approaches to "missingness" and demonstrate their advantages and disadvantages over discarding incomplete cases. With the FDA Amendments Act of 2007, drug and clinical trial data collection has been rapidly expanding, but these data are often sparse, and our dataset is no exception. Related studies (e.g., DiMasi et al. [70]) have typically used only complete case observations—discarding clinical trials with any missing information—which typically eliminates large portions of the data and may also lead to certain biases.

We use machine-learning techniques to form our predictions, including cross-validation for training and a held-out testing set for performance evaluation, and use the standard "area under the receiver operating characteristic curve" (AUC) metric to measure model performance. (AUC is the estimated probability that a classifier will rank a positive outcome higher than a negative outcome [39].) We achieve AUCs of 0.78 for predicting phase 2 to approval (95% confidence interval (CI): [0.75, 0.81]) and 0.81 for predicting phase 3 to approval (95% CI: [0.78, 0.83]). A time-series, walk-forward analysis approach shows similar results. We also apply our models to the current drug pipeline—that is, all drugs still in development as of the end of our dataset—to identify the candidates that have the highest and lowest probabili-

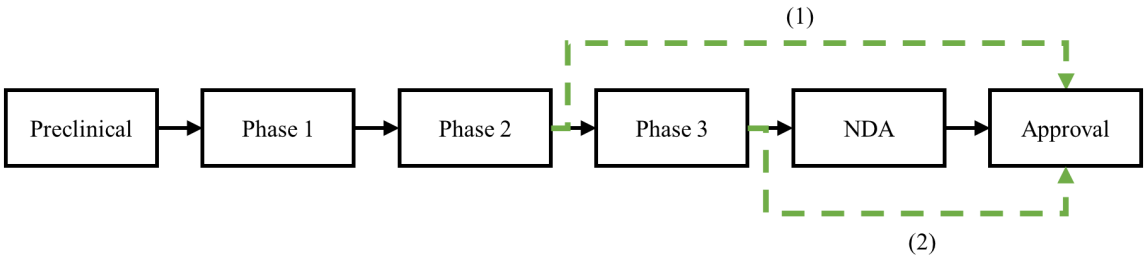


Figure 3-1: Predictive models for assessing the probability of approval of drug candidates in two scenarios: (1) after phase 2 testing and (2) after phase 3 testing.

ties of success. We examine the latest development statuses of these pipeline drug-indication pairs—a true "out-of-sample" experiment (validation on data not used in model building)—and find that candidates with higher scores are, indeed, more likely to progress to later clinical stages. This indicates that our classifiers do discriminate between high- and low-potential candidates.

3.2 Data

3.2.1 Summary Statistics

The commercial data vendor Informa® offers two databases that are used in our analysis: Pharmaprojects, which specializes in drug information, and Trialtrove, which specializes in clinical trials information [65]. These two databases aggregate drug and trial information from over 30,000 data sources in more than 150 countries, including company press releases, government drug databases (e.g., Drugs@FDA) and trial databases (e.g., clinicaltrials.gov [71] and clinicaltrialsregister.eu), and scientific conferences and publications. Using these sources, we construct two datasets of drug-indication pairs: phase 2 to approval (P2APP) and phase 3 to approval (P3APP). We extract clinical trial features from Trialtrove, and augment this data using drug features from Pharmaprojects. Applying machine-learning algorithms to these datasets allows us to estimate: (1) whether a drug-indication pair that has concluded phase 2 testing will be approved eventually; and (2) whether a pair that has concluded phase 3 testing will be approved eventually. Data cleaning procedures are outlined in

Appendix B.1.

We consider all indications associated with a particular drug, as opposed to only the lead indication. We extract all features that could conceivably correlate with the likelihood of success, from drug compound attributes (31 features from Pharmaprojects profiles) to clinical trial characteristics (113 features from Trialtrove). These features are defined in Table 3.1 and Appendix B.1. In general, each dataset may be partitioned into two disjoint subsets: one with samples that have known outcomes, and another with samples that are still in the pipeline at the time of snapshot of the databases (that is, the outcomes are unknown). To provide intuition for the characteristics of the samples, we describe key summary statistics of each subset.

The P2APP dataset consists of 6,344 drug-indication pairs that have ended phase 2 testing; that is, there are no phase 2 trials in progress or planned in the database. The phase 2 trials in this dataset range from August 8, 1990 to December 15, 2015. In our sample, 4,812 pairs have known outcomes, while 1,532 pairs are still in the pipeline. In the subset with known outcomes, we define the development statuses of suspension, termination, and lack of development as "failures" (86.8%), and registration and launch as "successes" or approvals (13.2%). The P3APP dataset consists of 1,870 pairs that have ended phase 3 testing, of which 1,610 pairs have known outcomes, while 260 pairs are still in the pipeline. For those pairs with known outcomes, we define "failures" (59.1%) and "successes" (40.9%) in the same fashion as the P2APP dataset. The phase 3 trials in P3APP span from January 1, 1988 to November 1, 2015. These figures are summarized in Table 3.2. Here, the use of terms "success" and "failure" is in the context of achieving approval. We note that our definition of "failures" can include drug development programs that are terminated due to factors unrelated to the performance of the drug (e.g., market conditions, business decisions). In Section 3.4, we find that this outcome variable has significant associations with trial performance and other factors.

The datasets cover 15 indication groups: alimentary, anti-infective, anti-parasitic, blood and clotting, cardiovascular, dermatological, genitourinary, hormonal, immunological, musculoskeletal, neurological, anti-cancer, rare diseases, respiratory, and sen-

Table 3.1: Description of features extracted from Pharmaprojects and Trialtrove. Some parent features are multi-label. We transform all multi-label parent features into binary child features. Drug-indication pairs for the same drug have the same drug features; drug-indication pairs involved in the same trial have the same trial features.

	Description	**Type**
Drug Features		
Route	Route of administration of the drug, the path by which the drug is taken into the body.	Multi-label
Origin	Origin of the active ingredient in the drug.	Multi-label
Medium	Medium of the drug.	Multi-label
Biological target family	Family of proteins in the body whose activity is modified by the drug, resulting in a specific effect.	Multi-label
Pharmacological target family	Mechanism of action of the drug, the biochemical interaction through which the drug produces its pharmacological effect.	Multi-label
Drug-indication development status	Current phase of development of the drug for the indication.	Binary
Prior approval of drug for another indication	Approval of the drug for another indication prior to the indication under consideration (specific to drug-indication pair).	Binary
Trial Features		
Duration	Duration of the trial (from reported start date to end date) in days.	Continuous
Study design	Design of the trial (keywords).	Multi-label
Sponsor type	Sponsors of the trial grouped by types.	Multi-label
Therapeutic area	Therapeutic areas targeted by the trial.	Multi-label
Trial status	Status of the trial.	Binary
Trial outcomes	Results of the trial.	Multi-label
Target accrual	Target accrual of the trial.	Continuous
Actual accrual	Actual accrual of the trial.	Continuous
Locations	Locations of the trial by country.	Multi-label
Number of identified sites	Number of sites where the trial was conducted.	Continuous
Biomarker involvement	Type of biomarker involvement in the trial.	Multi-label
Sponsor track record	Sponsor's success in developing other drugs prior to the drug-indication pair under consideration.	Continuous
Investigator experience	Primary investigator's success in developing other drugs prior to the drug-indication pair under consideration.	Continuous

sory products. Anti-cancer agents make up the largest subgroup in P2APP, and the second largest in P3APP (see Table 3.3). Industry-sponsored trials dominate both datasets (see Table 3.4). In aggregate, we observe a decreasing trend in success rates over five-year rolling windows from 2003 to 2015 (see Fig. 3-2).

To the best of our knowledge, this sample is the largest of its kind. All prior published research in this literature involved fewer than 100 drugs or 500 trials [66, 67, 68, 70]. In addition, our datasets cover a diverse set of indication groups, as opposed to a single area such as oncology.

Table 3.2: Sample sizes of P2APP and P3APP datasets. We consider phase 2 trial information in P2APP datasets and phase 3 trial information in P3APP dataset.

	Counts				
	Drug-indication Pairs	**Phase 2/3 Trials**	**Unique Drugs**	**Unique Indications**	**Unique Phase 2/3 Trials**
			P2APP		
Success	635	2,563	540	173	2,486
Failure	4,177	10,328	2,779	263	9,722
Pipeline	1,532	2,815	1,189	221	2,713
Total	6,344	15,706	4,073	288	14,584
			P3APP		
Success	659	1,830	572	171	1,801
Failure	951	2,425	764	203	2,360
Pipeline	260	494	240	120	480
Total	1,870	4,749	1,451	253	4,552

Table 3.3: Breakdown of drug-indication pairs by indication groups. A drug-indication pair may have multiple indication group tags. For instance, renal cancer is tagged as both anti-cancer and rare disease.

	Counts	
	P2APP	**P3APP**
All	6,344	1,870
Anti-cancer	2,239	409
Rare Diseases	1,105	259
Neurological	1,069	444
Alimentary	757	249
Immunological	474	101
Anti-infective	493	177
Respiratory	428	134
Musculoskeletal	394	121
Cardiovascular	388	158
Dermatological	254	45
Genitourinary	210	85
Blood and Clotting	160	97
Sensory	137	41
Hormonal	17	4
Anti-parasitic	8	0

Table 3.4: Breakdown of trials by sponsor types. A trial may be sponsored by more than one party (e.g., collaboration between industry developers and academia).

	Counts	
	P2APP	**P3APP**
All	14,584	4,552
Other Pharma	5,432	1,721
Top 20 Pharma	5,322	2,369
Academic	4,869	736
Government	1,807	314
Cooperative Group	958	230
Not for Profit	181	51
Generic	52	54
Contract Research Organization	41	17

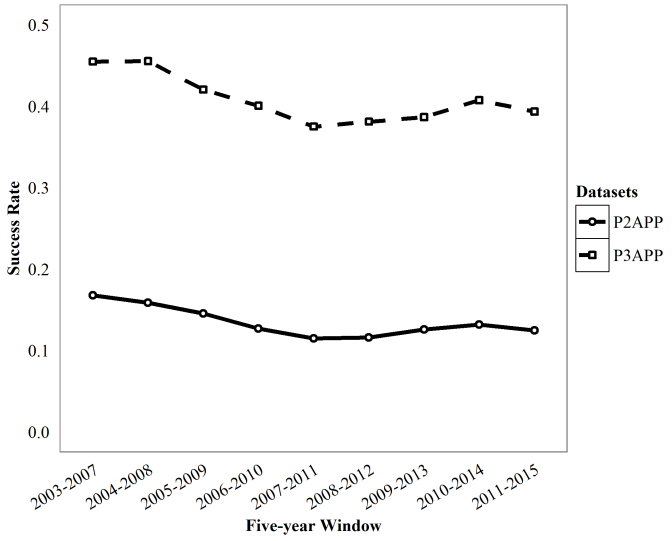


Figure 3-2: Success rates in P2APP and P3APP over five-year rolling windows from 2003–2015.

3.2.2 Missing Data

Prior to the 2007 FDA Amendments Act (FDAAA), it was not uncommon for investigators to release only partial information about pipeline drugs and clinical trials to protect trade secrets or simply because there was no incentive to do more. Even today, some investigators still do not adhere to the FDAAA-mandated registration policy or submit adequate registrations. Therefore, all historical drug development databases have missing data. We note that the "missingness" here is largely related to the post-study reporting of clinical trial data as opposed to in-trial data missingness (e.g., censorship of panel data due to patients terminating trial participation prematurely). In the former case, the data (e.g., trial duration, trial outcomes) are usually available to the investigators but may not be released publicly, and are thus considered "missing" from our standpoint.

Figs. 3-3 and 3-4 and Tables 3.5 and 3.6 summarize the patterns of missingness in our dataset (we exclude pipeline drug-indication pairs here because their outcomes are still pending). The missing data patterns are multivariate. When conditioned on the latest level of development, for any indication, we find that successful drugs generally have lower levels of missingness compared to failed drugs. For instance, in the P2APP dataset, 61% of failed drugs have an unknown medium, while only 15% of approved drugs are missing this feature. We also observe that completed trials tend to have greater levels of missingness than terminated trials. Between two datasets, we find that the P3APP dataset, which focuses on phase 3 drugs and trials, generally has less missing data for both drug and trial features than the P2APP dataset which focuses on phase 2 drugs and trials. This is expected since phase 3 trials are primarily used to support registration filings.

Most related studies do not report the extent of missing data in their samples, presumably because smaller datasets were used. DiMasi et al. [70] reported missing data for some of their factors, and addressed it through listwise deletion—deleting all observations with any missing factors. Since statistical estimators often require complete data, this approach is the simplest remedy for missingness. However, it

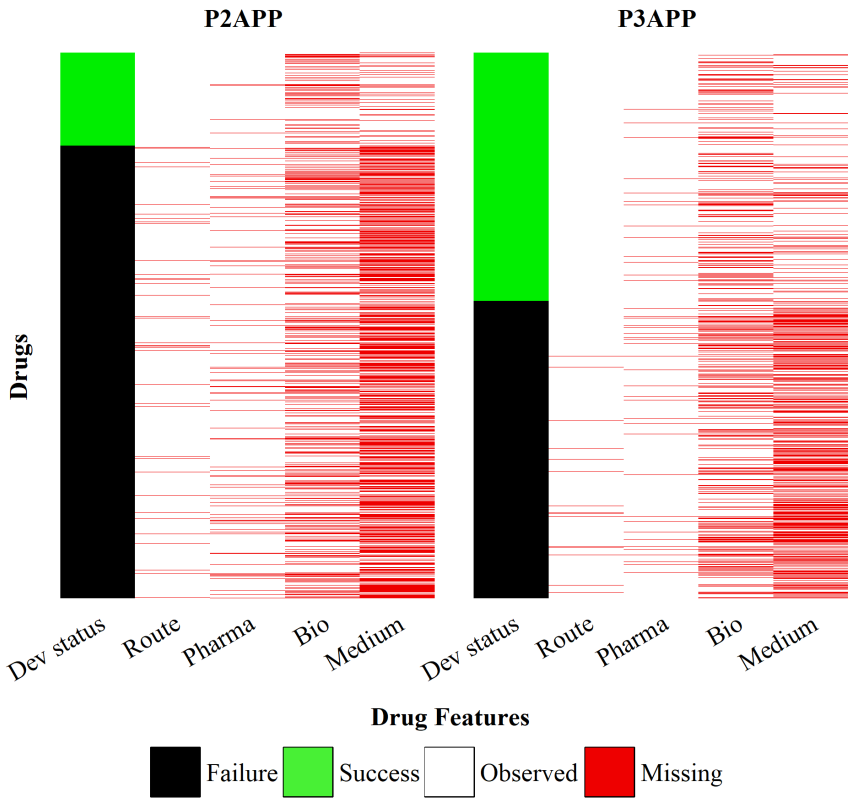


Figure 3-3: Missingness patterns of drug features. Each row corresponds to a unique drug. Features not included in the figure are complete and do not have missing values. *Abbreviations*: dev status, highest level of development of a drug for any indication; pharma, pharmacological target family; bio, biological target family.

greatly reduces the amount of data available and decreases the statistical power of the resulting statistics. Furthermore, listwise deletion is valid only under strict and unrealistic assumptions (see Section 3.3.1), and when such conditions are violated, inferences are biased. In this study, we make an effort to include in our analysis all observed examples, with or without complete features, through the use of statistical imputation.

3.3 Methods

Our analysis consists of two parts. First, we impute missing values to generate complete datasets. Next, we apply a range of machine-learning algorithms to build predictive models based on the imputed data. Illustration of the specific components of

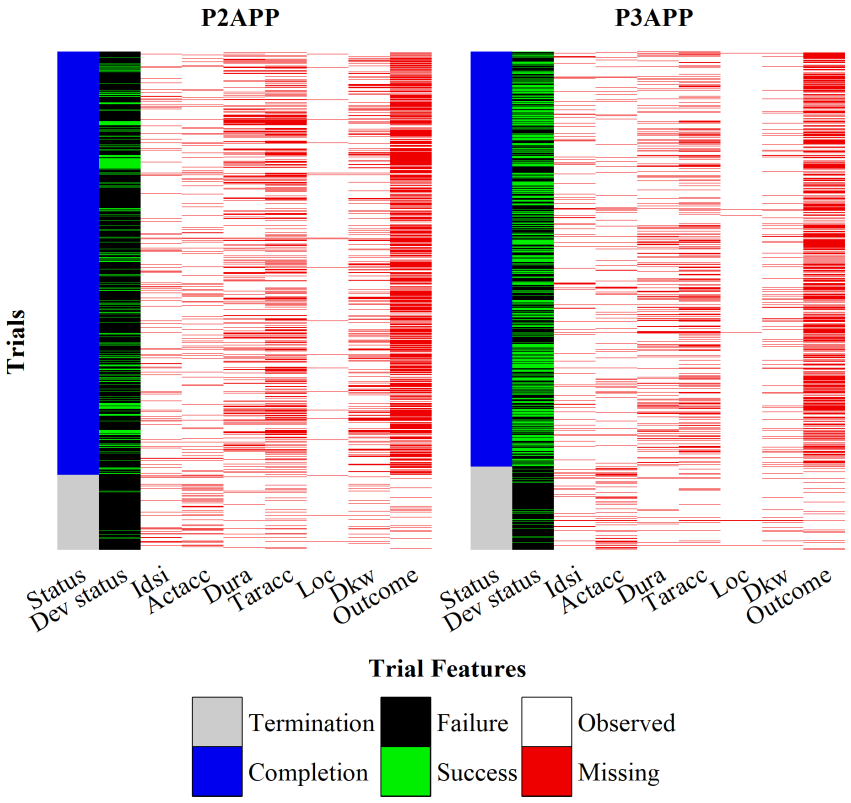


Figure 3-4: Missingness patterns of trial features. Each row corresponds to a unique clinical trial. Features not included in the figure are complete and do not have missing values. *Abbreviations*: dev status, highest level of development of a drug for any indication; status, trial status; idsi, number of identified sites; actacc, actual accrual; dura, duration; taracc, target accrual; loc, locations; dkw, trial study design keywords; outcome, trial outcomes.

Table 3.5: Missingness in drug features with respect to unique drugs.

	Missingness		
	Unconditional	**Success**	**Failure**
P2APP			
Route	0.04	0.00	0.04
Pharmacological target family	0.06	0.02	0.07
Biological target family	0.32	0.27	0.32
Medium	0.53	0.15	0.61
P3APP			
Route	0.01	0.00	0.02
Pharmacological target family	0.03	0.02	0.04
Biological target family	0.27	0.24	0.30
Medium	0.35	0.14	0.54

Table 3.6: Missingness in trial features with respect to unique trials.

	Missingness		
	Unconditional	**Completion**	**Termination**
	P2APP		
Number of identified sites	0.10	0.10	0.10
Actual accrual	0.12	0.10	0.22
Duration	0.26	0.29	0.05
Target accrual	0.37	0.42	0.09
Locations	0.02	0.02	0.02
Study design keywords	0.22	0.24	0.10
Trial outcomes	0.63	0.73	0.11
	P3APP		
Number of identified sites	0.10	0.09	0.12
Actual accrual	0.12	0.09	0.26
Duration	0.17	0.19	0.06
Target accrual	0.27	0.31	0.09
Locations	0.01	0.01	0.02
Study design keywords	0.09	0.09	0.06
Trial outcomes	0.53	0.62	0.07

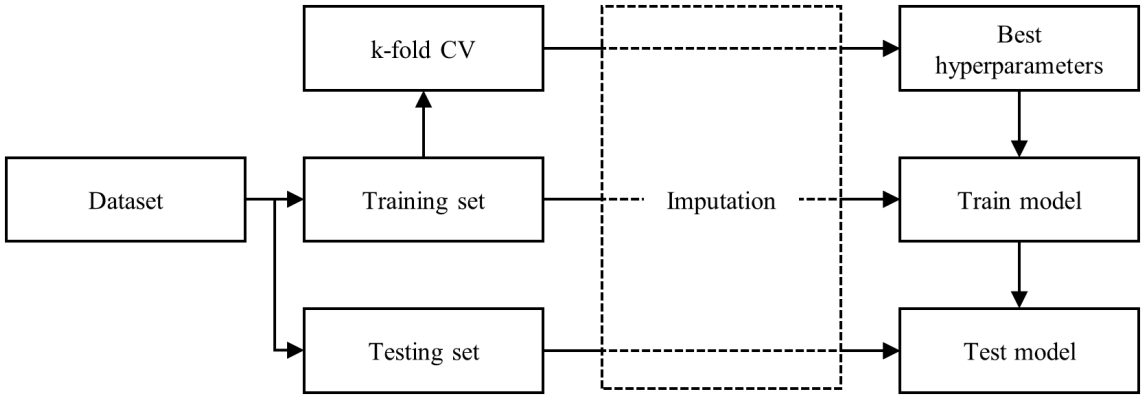


Figure 3-5: Modeling methodology adopted in this study. *Abbreviations*: CV, cross-validation.

our analysis appear in Fig. 3-5.

3.3.1 Statistical Imputation

Missing data may be classified into three categories [72]: missing completely at random (MCAR), missing at random (MAR), and missing not-at-random (MNAR). MCAR refers to data that are missing for reasons entirely independent of the data; MAR applies when the missingness can be fully accounted for by the observed variables; and MNAR refers to situations when neither MCAR nor MAR is appropriate,

in which case the probability of missingness is dependent on the value of an unobserved variable [73].

For a more precise definition, let Y denote a n by p data matrix (with elements y_{ij}) where the rows represent samples and the columns represent variables. We further partition the observed part of Y as Y_{obs} and the missing part of Y_{mis}, so collectively:

$$Y = (Y_{obs}, Y_{mis}) \tag{3.1}$$

Next, let R be a n by p response indicator matrix where elements $r_{ij} = 0$ if the corresponding element y_{ij} is missing and $r_{ij} = 1$ if y_{ij} is observed. The distribution of R, known as the missing data model/missingness mechanism, may be written generally as $P(R \mid Y_{obs}, Y_{mis}, \xi)$, where ξ parameterizes the relationship between R and Y. The missingness is said to be MCAR if the probability of missingness is totally unrelated to the data:

$$P(R \mid Y_{obs}, Y_{mis}, \xi) = P(R \mid \xi) \tag{3.2}$$

The missingness is said to be MAR if the missingness is independent of the values of the missing data when conditioned on the observed data:

$$P(R \mid Y_{obs}, Y_{mis}, \xi) = P(R \mid Y_{obs}, \xi) \tag{3.3}$$

Finally, the missingness is said to be MNAR if the probability of missingness, $P(R \mid Y_{obs}, Y_{mis}, \xi)$, cannot be simplified, i.e., the probability of missingness depends on the unobserved underlying values of the missing data and/or of other observed variables.

Now, we let the distribution of Y, which is the data model we are interested in, be described by some parameters θ. The missingness mechanism can be further described as ignorable under two conditions. First, the missingness must be MAR. Second, the parameters θ and ξ must be distinct, meaning θ and ξ should be a priori independent where $P(\theta, \xi)$ factors into $P(\theta)P(\xi)$ [74]. In many situations, the second

condition is reasonable because knowing θ provides little information about ξ and vice versa [75]. In general, the first requirement of MAR is considered to be the more important condition. When ignorability holds, Rubin [76] showed that:

$$P(Y_{mis} \mid Y_{obs}, R) = P(Y_{mis} \mid Y_{obs}) \tag{3.4}$$

This implies that the distribution of the data is independent of the missing data model, and is identical in both the observed and unobserved groups [73]:

$$P(Y \mid Y_{obs}, R = 1) = P(Y \mid Y_{obs}, R = 0) \tag{3.5}$$

In this case, we can model the conditional distribution $P(Y \mid Y_{obs}, R = 1)$ from the observed data, and use it to draw imputations for the missing data. In other words, the missing data model R can be ignored and not modeled. If the missingness is nonignorable, then Eq. (3.5) does not hold, and the distributions are not equivalent. When this happens, we need to estimate the missingness mechanism, and incorporate it into the imputation model.

If the missingness is MCAR, the observed samples can be viewed as a random subsample of the dataset. Consequently, using listwise deletion should not introduce any bias. While convenient, this assumption is rarely satisfied in practice. In most drug-development databases, failed drugs are more likely to have missing features than successful drugs (see Table 3.5). Clearly, MCAR does not hold.

Applying listwise deletion when the missingness is not MCAR can lead to severely biased estimates. Moreover, given the nature of drug-development reporting, a large portion of the original data may be discarded if many variables have missing values. For these reasons, the listwise-deletion approach adopted by DiMasi et al. [70] and others is less than ideal.

Given only the observed data, it is impossible to test for MAR versus MNAR [77]. However, our knowledge of the data-collection process suggests that MAR is a plausible starting point, and we hypothesize that the missingness in drug and trial features is mainly accounted for by drug development and trial statuses, respectively.

Our observations in Tables 3.5 and 3.6 support this approach, as the missingness proportions for some features differ greatly depending on the outcome.

Our assumption of MAR is consistent with the data-collection methodology in the Informa® databases. Drug profiles are built up over time in Pharmaprojects. As a drug advances to later phases, information about its characteristics becomes more readily available because investigators release more data about pipeline drugs after each phase of clinical testing. Informa® inputs this information into its databases as they become available in the public domain or through primary research. Approved drugs are more likely to have more complete profiles, while information about failed drugs tends to stay stagnant because no further studies are conducted. It is very plausible that the MAR nature of our datasets is an artifact of data collection, and by extension, so are similar pharmaceutical datasets extracted from the public domain and maintained in the same fashion. Originally intended to track drug and trial activities, Pharmaprojects and Trialtrove are not structured to keep track of information updates over time since there was no use for it. Without timestamps of the updates, we are not able to eliminate the MAR artifact from our datasets.

In our analysis, we impute the missing data under the more plausible MAR assumption to obtain complete datasets. In contrast to listwise deletion, we fill in missing values using information in the observed variables. This allows us to utilize data that would otherwise be discarded. Thereafter, we can apply all the usual statistical estimators to this imputation-completed data.

We explore complete cases analysis and four imputation techniques commonly used in social science research and biostatistics: unconditional mean imputation, k nearest neighbors imputation, multiple imputation, and decision tree algorithms.

Complete Cases Analysis

In complete cases analysis (also known as listwise deletion), we discard all observations with missing data, in which case there is no imputation. It is the default method in many statistical programs. This method is generally not recommended because it is valid only under strict MCAR conditions, which rarely holds in practice.

Applying this approach to MAR/MNAR data will likely yield biased inferences. It is apparent that the dataset under study is not MCAR. Nevertheless, we can use this as comparison against other methods.

Unconditional Mean Imputation

In unconditional mean imputation, we fill in the missing values of a variable with the mean/mode (for continuous/nominal variables, respectively) of the observed cases of that variable. This method is also highly discouraged because it distorts the data distribution by reducing variability and undermining relationships between variables. The use of mean imputation is non-ideal, nevertheless it can be used as a baseline. In this study, we implement two variants: mean/mode and median/mode imputation.

k Nearest Neighbors Imputation

In k nearest neighbors imputation (kNN), given an instance with missing values, we select the k most similar cases that do not have missing values in the features to be imputed. As the name suggests, the replacements for the missing values are chosen from these k nearest neighbors. In this study, we use the Gower distance as a measure of similarity: the range-normalized Manhattan distance for continuous variables and the Jaccard distance for categorical variables. We explore five and ten nearest neighbors. For each missing value to be imputed, we use the median and mode, for continuous variables and binary variables, respectively, of the corresponding feature of the k closest neighbors as imputation.

Multiple Imputation

Multiple imputation (MI) is a principled missing data method that involves three steps: imputation, analysis, and pooling. In the first step, we specify an imputation model for each incomplete variable in the form of a conditional distribution, that is, missing data conditioned on the observed data. Then we draw multiple plausible values for each missing data point according to the specified variable models, creating multiple imputed datasets from one incomplete dataset. In this study, we specify

linear regression models for continuous variables and logistic regression models for categorical variables. In the second step, we analyze each imputed dataset individually using standard statistical procedures. Finally, in the third step, we pool the estimates obtained from the multiple individual analyses (e.g., probability predictions and regression coefficients) using Rubin's rules [76] to yield a single estimate. See Appendix B.2 for more details.

Decision Tree Algorithms

Decision trees are commonly used as predictive models. In contrast to most machine-learning algorithms, some decision tree algorithms can handle missing values internally without the need for imputation. In this study, we focus on the C5.0 algorithm. C5.0 is a tree-based model developed by Quinlan [78]. It uses entropy as the node impurity measure. When considering a variable for a split, C5.0 uses only examples for which that variable is not missing to calculate the node impurity. When an instance sent down C5.0 encounters a split variable for which it has a missing value, it is split into the branches fractionally, according to the split proportion of the observed instances.

3.3.2 Machine Learning Models

We formulate our two scenarios as supervised bipartite ranking problems, where the goal is to predict the outcome—success or failure—of a drug-indication pair given a set of input features. Initially, we split each dataset into training and testing sets. For each scenario, we train various classifiers based on the corresponding training set, and compute the expected error of our predictive models by testing them on the held-out testing set.

We create feature matrices from the datasets by representing drug and trial features for each drug-indication pair as vectors (see Fig. 3-6). Drug-indication pairs associated with multiple trials are represented by the same number of feature vectors, e.g., a pair with two trials has two rows. We give a concrete example in Fig. 3-6.

Consider the drug-indication pair Analiptin-diabetes type 2 in the P2APP dataset. We represent it using two vector rows since it has two phase 2 trials in Trialtrove. Note that the feature matrix is incomplete due to missing drug and trial features. We also construct a column vector of labels, which contains the outcomes of the drug-indication pairs. Labels are not available for pipeline drug-indication pairs because they are still in development and their outcomes are still uncertain, hence these observations are not used to train our classifiers. However, with the trained classifiers, we can generate predictions for pipeline data.

We split each dataset (excluding pipeline drugs-indication pairs) into two disjoint sets, one training set and one testing set, and form feature matrices for both according to the drug-indication pairs in each set. The testing sets serve as out-of-sample datasets to evaluate our models. Therefore, we mask their outcomes (that is, we treat them as unknown) and access them only at the very end to check our performance.

To deal with missing data in both training and testing sets, we consider the imputation techniques described in Section 3.3.1. We follow best practices of the missing-data literature by including as many relevant auxiliary variables as possible, as well as all variables used in subsequent models [77, 79, 80, 81]. This makes the assumption of MAR more plausible in our datasets, and helps to reduce bias in subsequent analyses [75]. In particular, it is necessary to include our target variable—the drug-indication development status—in our imputation model because we hypothesize that missingness is mainly accounted for by it. This is not an issue for the training sets. However, the outcomes in the testing sets are masked, and not supposed to be known. Therefore, we treat the testing set outcomes as though they were missing and impute them together with all the other missing features. After imputation, we discard the imputed testing-set outcomes, and use only the imputed feature values for predictions. We do the same when evaluating pipeline datasets.

With respect to the machine-learning algorithm, we explore several linear and non-linear classifiers commonly used in literature, including penalized logistic regression (PLR), random forests (RF), neural networks (NN), gradient boosting trees (GBT), support vector machines with radial basis functions (SVM), and decision trees C5.0.

We implement the first five algorithms in Python [38] and the last in R [82].

For training, we weight each feature matrix row example according to the number of trials of the corresponding drug-indication pair. In our earlier example, the drug-indication pair Analiptin-diabetes type 2 was involved in two phase 2 trials. It is represented by two vector rows in the feature matrix (see Fig. 3-6). Both rows are used as training examples, and each is weighted equally during training (0.5, since there are two trials in total). To obtain predictions for a drug-indication pair, we average the output probabilities and scores of the corresponding feature vector rows that are used as inputs to the classifier.

All machine-learning algorithms have hyper-parameters that affect the flexibility of the model and must be tuned to each dataset to optimize goodness of fit. Poorly-chosen hyper-parameters can lead to overfitting (attributing signal to noise) or underfitting (attributing noise to signal). We tune our parameters using k-fold cross-validation (with $k = 5$ or $k = 10$, depending on the sample size). Since the cross-validation process should emulate the testing process as closely as possible, we include imputation in the cross-validation loop as well. We split the training set into validation and non-validation folds. Then we treat validation fold outcomes as missing, and impute them as we would for a testing set. From here, we ignore the imputed validation fold outcomes and proceed with the standard validation process.

In the final step, we test the trained classifiers on the unseen testing sets for out-of-sample model validation. This gives the expected performance of our predictive models for each of the scenarios, using the standard AUC metric to measure model performance.

3.4 Results

3.4.1 Imputation Versus Listwise Deletion

We study the effects of imputation using a "gold-standard" dataset derived from the complete cases of the P2APP dataset (see Table 3.7). To simulate the missingness

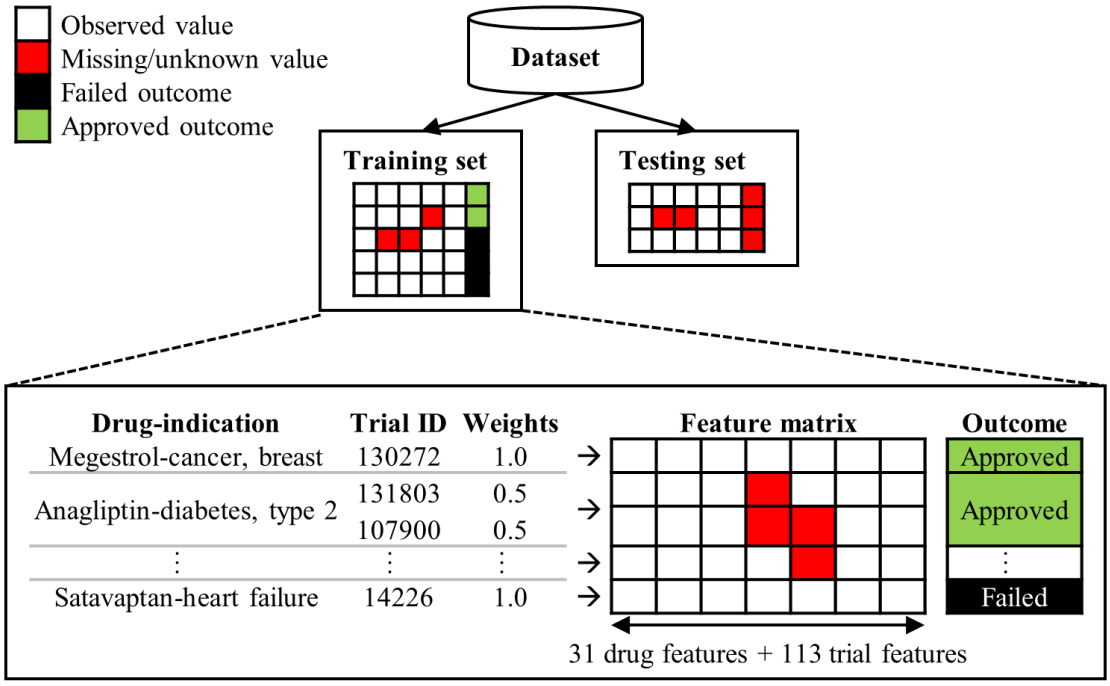


Figure 3-6: Feature matrix of dataset. Each row corresponds to a feature vector; each feature corresponds to an entry in the vector; each vector has a length of 144 since we have 31 drug and 113 trial features. Feature vectors of all drug-indication pairs in the dataset form the feature matrix collectively. Trial ID is a unique trial identifier in Trialtrove.

present in the original dataset, we introduce missingness in the gold-standard dataset based on our MAR assumption and the missingness patterns observed in the P2APP dataset. We randomly split the drug-indication pairs into a training set (70%) and a testing set (30%), and use five different missing data approaches, as described in Section 3.3.1, to generate complete training sets from the MAR training set. We use each imputed training set to build six different predictive models (PLR, RF, NN, GBT, SVM, and C5.0) according to the methodology outlined in Section 3.3.2. We repeat this experiment 100 times for robustness. Table 3.8 summarize the AUC performance of the classifiers on the gold-standard testing sets. See Appendix B.3 for a more detailed description and results.

For all six machine-learning algorithms, we find that gold-standard classifiers—that is, the models derived from complete data—consistently outperform their complete case analysis and imputation counterparts. This is logical because useful information is invariably lost when we introduce missingness in the datasets. In contrast, complete case analysis often leads to inferior performance. The AUCs of classifiers trained on complete cases training sets tend to be smaller than those trained on imputed training sets. This suggests that imputation does indeed offer improved fit and predictive power over listwise deletion.

Overall, we find kNN imputation to be most compatible with our datasets. It provides the least biased imputations among all missing data methods (see Appendix B.3). In particular, the combination of kNN imputation ($k = 5$) with RF gives one of the highest gold-standard testing set AUCs (0.81). We note a few other MI combinations that yield comparable or marginally better performance but focus on the 5NN-RF approach in subsequent analyses on the main datasets due to its ease of implementation and application. We find that SVM has the worst performance among all machine-learning models. This is not surprising because SVMs are aimed only at learning binary classifiers, and do not generally produce good class probability estimates. Consequently, such models do not necessarily give high AUCs.

We also compare our approach with the ANDI algorithm [70] by applying a modified version of the index on oncology drugs in the gold-standard testing sets (see

Table 3.7: Sample size of the gold-standard dataset (derived from complete cases of P2APP).

	Counts				
	Drug-indication Pairs	**Phase 2 Trials**	**Unique Drugs**	**Unique Indications**	**Unique Phase 2 Trials**
Success	166	341	152	83	337
Failure	812	1,672	503	158	1,549
Total	978	2,013	623	171	1,872

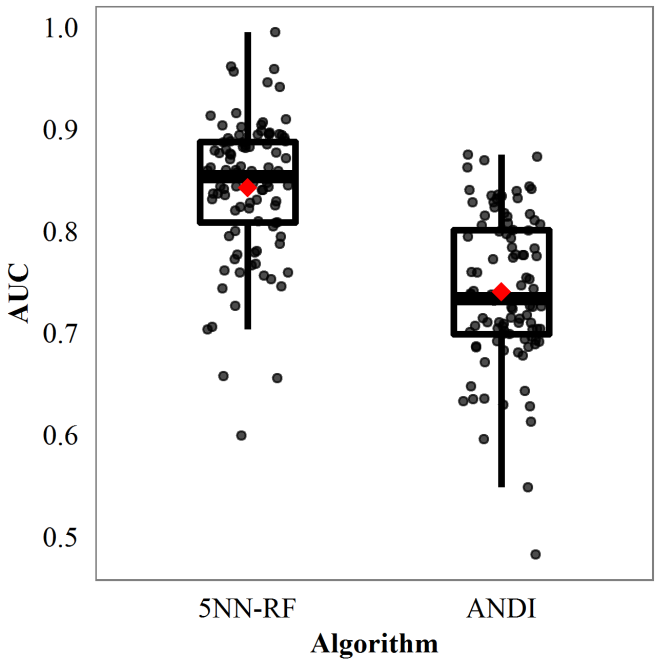


Figure 3-7: Distributions of AUC of 5NN-RF and the modified ANDI on oncology-only gold-standard testing sets.

Appendix B.4 for a more in-depth description). We find that our 5NN-RF model achieves significantly higher AUC than the modified ANDI, with an average improvement of 0.1 in AUC over 100 simulations (see Fig 7). We believe that this gain can be attributed to a larger training set with a wider range of features, a nonlinear model that can capture the complex relationships in the data, and a proper model validation methodology.

Table 3.8: Out-of-sample performance of different missing data approaches. *Abbreviations*: avg, average; sd, standard deviation; 5%, 5th percentile; 50%, median; 95%, 95th percentile; m, number of imputations generated.

	Testing Set AUC				
	Avg	**Sd**	**5%**	**50%**	**95%**
			PLR		
Gold Standard	0.810	0.028	0.761	0.808	0.853
Complete Cases	0.755	0.040	0.683	0.764	0.813
Mean/mode	0.778	0.031	0.729	0.779	0.823
Median/mode	0.778	0.031	0.728	0.779	0.824
5NN	0.786	0.032	0.738	0.787	0.834
10NN	0.787	0.032	0.739	0.791	0.835
MI ($m = 1$)	0.781	0.036	0.722	0.777	0.843
MI ($m = 10$)	0.782	0.031	0.729	0.782	0.831
			RF		
Gold Standard	0.837	0.027	0.793	0.837	0.876
Complete Cases	0.764	0.048	0.685	0.772	0.830
Mean/mode	0.775	0.031	0.726	0.771	0.822
Median/mode	0.774	0.031	0.723	0.774	0.827
5NN	0.805	0.033	0.755	0.805	0.857
10NN	0.802	0.033	0.747	0.805	0.856
MI ($m = 1$)	0.797	0.033	0.748	0.795	0.853
MI ($m = 10$)	0.804	0.030	0.751	0.804	0.848
			NN		
Gold Standard	0.800	0.032	0.754	0.799	0.849
Complete Cases	0.715	0.043	0.638	0.716	0.779
Mean/mode	0.790	0.037	0.739	0.789	0.848
Median/mode	0.789	0.036	0.740	0.792	0.849
5NN	0.794	0.032	0.743	0.798	0.842
10NN	0.797	0.036	0.737	0.798	0.851
MI ($m = 1$)	0.780	0.036	0.719	0.781	0.838
MI ($m = 10$)	0.795	0.030	0.750	0.795	0.838

Table 3.8 (continued): Out-of-sample performance of different missing data approaches. *Abbreviations*: avg, average; sd, standard deviation; 5%, 5th percentile; 50%, median; 95%, 95th percentile; m, number of imputations generated.

	Testing Set AUC				
	Avg	**Sd**	**5%**	**50%**	**95%**
			GBT		
Gold Standard	0.820	0.028	0.776	0.821	0.868
Complete Cases	0.746	0.050	0.659	0.756	0.816
Mean/mode	0.781	0.034	0.724	0.784	0.826
Median/mode	0.778	0.033	0.719	0.783	0.823
5NN	0.796	0.029	0.737	0.798	0.837
10NN	0.796	0.028	0.748	0.798	0.838
MI ($m = 1$)	0.796	0.031	0.747	0.796	0.847
MI ($m = 10$)	0.804	0.031	0.757	0.803	0.854
			SVM		
Gold Standard	0.785	0.030	0.730	0.786	0.831
Complete Cases	0.733	0.053	0.650	0.741	0.795
Mean/mode	0.766	0.036	0.707	0.771	0.818
Median/mode	0.764	0.035	0.711	0.771	0.818
5NN	0.771	0.034	0.722	0.770	0.827
10NN	0.772	0.037	0.710	0.773	0.825
MI ($m = 1$)	0.760	0.035	0.696	0.762	0.813
MI ($m = 10$)	0.768	0.030	0.719	0.764	0.813
			C5.0		
Gold Standard	0.800	0.033	0.758	0.800	0.844
Complete Cases	0.710	0.063	0.585	0.713	0.802
Mean/mode	0.758	0.039	0.698	0.762	0.816
Median/mode	0.754	0.043	0.679	0.751	0.823
5NN	0.772	0.038	0.715	0.772	0.843
10NN	0.770	0.035	0.710	0.771	0.822
MI ($m = 1$)	0.758	0.037	0.701	0.754	0.819
MI ($m = 10$)	0.807	0.031	0.756	0.808	0.857

3.4.2 Predicting Drug Approvals

We analyze the two main datasets (P2APP and P3APP) by first splitting each into a training set (70%) and a testing set (30%) randomly (pipeline drug-indication pairs are omitted since their outcomes have yet to be determined). Subsequently, we train 5NN-RF models for each scenario. We repeat this experiment 100 times for robustness. Table 3.9 summarizes the AUC performance metrics for the testing sets. On average, we achieve 0.78 AUC for P2APP and 0.81 AUC for P3APP.

The observed performance is essentially the MAR testing set AUC, since backfilling has already affected the datasets used. In Appendix B.3, we highlight the perils of relying on the MAR testing set for model validation, and suggest that the AUCs for the gold-standard and MCAR testing sets are more reflective of a classifier's real performance. Unfortunately, we have access to neither the gold-standard nor the MCAR testing sets, because we do not know the true, underlying values of the missing features. However, our experiments indicate that the AUCs for the MAR and MCAR testing sets of the 5NN-RF combination are very close (a difference of 0.002 on average). This means that we may use the former, the only observed figure, as a reasonable estimate of the latter, which reflects real performance.

Next, we train classifiers based on the union of the training and testing sets, and use them to generate predictions for pipeline drug-indication pairs. We generate predictions for P2APP using only information from phase 2 trials and for P3APP using only information from phase 3 trials. While we cannot compute AUC scores for these samples because their outcomes are still pending, we can compare their prediction scores with their development status at the time of this writing. These pipeline drug-indication pairs may still be in the same clinical stage (no change, i.e., phase 2 for P2APP; phase 3 for P3APP), be terminated (failed), or have progressed to higher phases (advanced).

Fig. 3-8 and Tables 3.10 and 3.11 summarize the distributions of pipeline prediction scores. We find that pairs that fail generally have lower scores than those that advance to later phases of development. In Fig. 3-8, we observe peaks at the lower end

of the score spectrum for failed pairs (red) for both datasets. In contrast, pairs that advance tend to have peaks at higher scores (green). We observe the same patterns when we disaggregate the distributions by indication groups: the green parts tend to cluster above the distribution median while the red parts cluster below. However, there are also some indication groups for which there are too few samples to make any useful remarks (e.g., hormonal products in P2APP). From Table 3.10, we see that the average scores of failed pairs are indeed lower than those that advance (differences ranging from 0.05 to 0.15). In Table 3.11, we bin drug-indication pairs that have new developments (whether failure or advancement) into four groupings, depending on their prediction scores. For each bin, we compute the proportion of samples that advance to later development stages. We find that the proportions generally increase with the score magnitude, suggesting that pairs with higher scores are more likely to advance than those with lower scores. We note that progress to later clinical stages does not always lead to approval. However, the results are still promising because advancement is a necessary condition for approval. Our experiments indicate that our trained classifiers are able to discriminate between high- and low-potential candidates.

To gain insight into the logic of our trained predictive models, we compute the average importance of features used in the 5NN-RF classifiers over all the experiments, and extract the top ten most informative variables. The RF classifier we used computes the importance of a variable by finding the decrease in node impurity for all nodes that split on that variable, weighted by the probability of reaching that node (as estimated by the proportion of samples reaching that node), averaged over all trees in the forest ensemble [38, 83]. Table 3.12 summarizes the results.

We find that trial outcome (whether the trial was completed with its primary endpoints met) and trial status (whether the trial was completed or terminated) have significant associations with success. These two features were consistently ranked the top two out of all variables and across both datasets. It is easy to imagine that a drug-indication pair whose trials were terminated has a low probability of success in terms of advancing from phase 2 or phase 3 to approval. In contrast, candidates that achieve positive outcomes certainly have a better shot at success. We also observe

that prior approval of a drug has an effect on success for new indications or patient segmentation. It is plausible that developing an approved drug for a new indication has a greater likelihood of success than a new candidate.

In addition, trial characteristics such as accrual, duration, and the number of identified sites frequently appear in the top ten important variables. There are several possible explanations. For example, trials that end quickly without achieving primary endpoints may undermine the likelihood of success, and drugs with trials that have small accrual—and thus low statistical power—may have a lower probability of being approved.

We also find sponsor track records—quantified by the number of past successful trials (trials that achieve positive results or meet primary endpoints)—to be a useful factor for prediction. This factor has not been considered in previous related studies, but the intuition for its predictive power is clear: strong track records are likely associated with greater expertise in drug development.

Since drugs developed for different indication groups may have very different characteristics, we might expect classifiers trained on indication-group-specific data to outperform general classifiers. We build and analyze such specialized classifiers by filtering the datasets by indication group before performing the experiment described in the previous section. As a comparison, we also break down the performance of the general classifiers by indication group. Table 3.9 shows the results for selected indication groups. In general, we find specialized models to give poorer performance than general models. This is likely because the former are trained on less data, which makes them less accurate and more susceptible to overfitting.

We note that the approach adopted in this section—splitting drug-indication pairs into training and testing sets randomly without considering the dates of development—may be less than ideal because of look-ahead bias. For example, if the results of a 2008 trial are included in the training set for predicting the outcome of a 2004 development path for a drug-indication pair, our model will be using future information during validation, which can yield misleading and impractical inferences. To address this issue, in the next section we apply our machine-learning framework to time-series

Table 3.9: Out-of-sample performance of classifiers. Comparison of the general and indication-group-specific classifiers. *Abbreviations*: avg, average; sd, standard deviation; 5%, 5th percentile; 50%, median; 95%, 95th percentile.

	Testing Set AUC									
	General					**Indication-Group-Specific**				
	Avg	**Sd**	**5%**	**50%**	**95%**	**Avg**	**Sd**	**5%**	**50%**	**95%**
P2APP										
All	0.777	0.017	0.749	0.775	0.806					
Anti-cancer	0.805	0.025	0.764	0.805	0.847	0.818	0.029	0.773	0.819	0.865
Rare Diseases	0.800	0.028	0.756	0.800	0.848	0.775	0.036	0.715	0.777	0.838
Neurological	0.767	0.036	0.710	0.769	0.819	0.778	0.039	0.721	0.779	0.834
Alimentary	0.749	0.045	0.672	0.751	0.817	0.732	0.048	0.651	0.734	0.807
Immunological	0.783	0.065	0.665	0.786	0.889	0.766	0.069	0.646	0.775	0.860
Anti-infective	0.735	0.043	0.673	0.736	0.800	0.750	0.047	0.684	0.746	0.832
Respiratory	0.756	0.055	0.648	0.764	0.835	0.867	0.043	0.794	0.872	0.921
Musculoskeletal	0.822	0.049	0.736	0.821	0.899	0.731	0.076	0.614	0.745	0.849
Cardiovascular	0.709	0.072	0.580	0.711	0.812	0.694	0.073	0.579	0.698	0.807
Genitourinary	0.633	0.086	0.503	0.634	0.790	0.706	0.091	0.552	0.710	0.840
P3APP										
All	0.810	0.018	0.781	0.810	0.834					
Anti-cancer	0.783	0.047	0.699	0.779	0.853	0.707	0.054	0.612	0.714	0.786
Rare Diseases	0.819	0.054	0.727	0.822	0.896	0.786	0.058	0.687	0.793	0.875
Neurological	0.796	0.037	0.734	0.794	0.857	0.789	0.038	0.741	0.787	0.853
Alimentary	0.817	0.047	0.744	0.820	0.891	0.805	0.054	0.718	0.808	0.888
Immunological	0.811	0.074	0.680	0.815	0.910	0.757	0.099	0.586	0.765	0.892
Anti-infective	0.757	0.065	0.644	0.752	0.854	0.708	0.068	0.600	0.707	0.808
Respiratory	0.823	0.065	0.712	0.831	0.920	0.773	0.083	0.627	0.784	0.907
Musculoskeletal	0.741	0.095	0.576	0.747	0.866	0.763	0.072	0.646	0.762	0.882
Cardiovascular	0.794	0.058	0.702	0.788	0.887	0.755	0.076	0.639	0.765	0.864
Genitourinary	0.814	0.083	0.670	0.821	0.937	0.801	0.090	0.635	0.808	0.927

data using rolling windows that account for temporal ordering in the construction of training and testing sets. Although this process makes use of less data within each estimation window than when the entire dataset is used, it minimizes the impact of look-ahead bias and yields more realistic inferences. We study the effects of random splitting versus temporal ordering in Appendix B.5.

3.4.3 Predictions Over Time

Drug development has changed substantially over time, thanks to new scientific discoveries and technological improvements. To reflect these changes in our predictive analytics, we adopt a time series, walk-forward approach to create training and testing sets for each of the two datasets, P2APP and P3APP (see Fig. 3-9). We sample five-year rolling windows between 2004 and 2014 from each dataset. Each window

Table 3.10: Distributions of prediction scores for all indication groups in aggregate. Advanced refers to progress to a higher phase from the original phase. Original phase for P2APP is phase 2; for P3APP is phase 3. For instance, out of 1,511 drug-indication pairs in the P2APP testing set, 859 pairs are still pending decision in phase 2, 244 pairs have failed, and 408 pairs have successfully advanced to phase 3 testing. *Abbreviations*: avg, average; sd, standard deviation; 5%, 5th percentile; 50%, median; 95%, 95th percentile; n, sample size.

	Prediction Scores					
	n	**Avg**	**Sd**	**5%**	**50%**	**95%**
			P2APP			
Aggregate	1,511	0.153	0.061	0.044	0.155	0.258
No change	859	0.143	0.060	0.041	0.147	0.246
Failed	244	0.137	0.061	0.034	0.147	0.240
Advanced	408	0.183	0.056	0.093	0.178	0.274
			P3APP			
Aggregate	252	0.417	0.189	0.128	0.402	0.695
No change	142	0.392	0.185	0.129	0.384	0.693
Failed	32	0.348	0.185	0.100	0.344	0.656
Advanced	78	0.492	0.176	0.233	0.492	0.699

Table 3.11: Distributions of prediction scores for all indication groups in aggregate. Proportion refers to the fraction of samples that advanced to a later phase from the original phase. *Abbreviations*: n, sample size.

Scores	**n**	**Proportion**
	P2APP	
< 0.1	108	0.231
0.1-0.2	368	0.671
0.2-0.3	171	0.766
≥ 0.3	5	1.000
	P3APP	
< 0.2	13	0.308
0.2-0.4	35	0.686
0.4-0.6	27	0.667
≥ 0.6	35	0.914

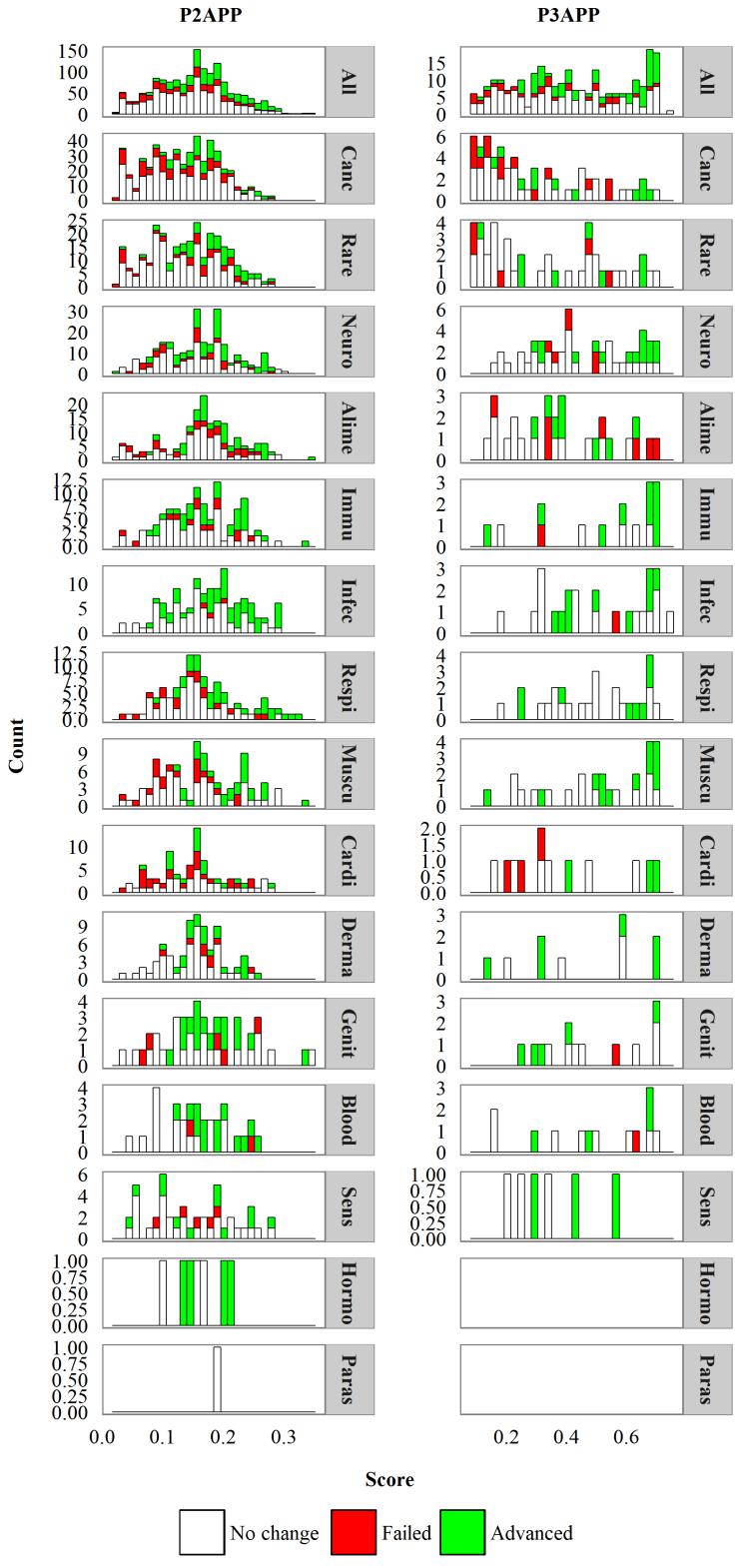


Figure 3-8: Distributions of prediction scores for P2APP and P3APP. First row for all indication groups in aggregate. Subsequent rows for specific indication groups.

Table 3.12: Top ten most important variables of 5NN-RF classifiers for P2APP and P3APP. Average and standard deviation taken across all experiments. *Abbreviations*: avg, average; sd, standard deviation.

	Importance	
	Avg	**Sd**
P2APP		
Trial outcome - completed, positive outcome or primary endpoint(s) met	0.234	0.043
Trial status	0.160	0.026
Medium - solution	0.051	0.018
Actual accrual	0.046	0.010
Sponsor type - industry, all other pharma	0.025	0.008
Sponsors track record - number of positive phase 3 trials	0.023	0.006
Sponsors track record - number of failed drug-indication pairs	0.021	0.007
Study design - placebo control	0.019	0.009
Target accrual	0.018	0.005
Prior approval of drug for another indication	0.018	0.007
P3APP		
Trial outcome - completed, positive outcome or primary endpoint(s) met	0.357	0.028
Trial status	0.148	0.014
Duration	0.099	0.016
Trial outcome - terminated, lack of efficacy	0.033	0.010
Trial outcome - completed, negative outcome or primary endpoint(s) not met	0.033	0.008
Therapeutic area - oncology	0.030	0.009
Prior approval of drug for another indication	0.021	0.007
Actual accrual	0.015	0.003
Medium - powder	0.014	0.007
Medium - solution	0.012	0.006

consists of a training set of drug-indication pairs whose outcomes become finalized within the window, and an out-of-sample, out-of-time testing set of drug-indication pairs that ended phase 2 or phase 3 testing, but are still in the pipeline with undetermined outcomes within the window. For example, consider the P2APP dataset. We draw the first window from 2004–2008, train our algorithm on drug-indication pairs that failed or were approved within this period as the training set, and apply the trained model to predict the outcomes of drug-indications that just ended phase 2 testing within the same window as the testing set.

We evaluate the resulting classifier by comparing its predictions with outcomes that are realized in the future (2009–2015). This rolling-window approach yields a total of eight overlapping training and testing periods where a new 5NN-RF model is trained for each period. The eighth testing period consists of drug-indication pairs in the pipeline at the time of snapshot of the databases. Unlike the first seven periods, their outcomes are still pending current development, and therefore we cannot

compute a testing AUC for this window. However, we can examine the predictions and compare the scores with their development statuses at the time of this writing.

Fig. 3-10 summarizes the results of the time-series analysis for the first seven windows. We observe an increasing trend over the years for both P2APP (0.67 in the first and 0.80 in the last window) and P3APP (0.77 in the first and 0.88 in the last window). Interestingly, we note that the proportions of complete cases in the training sets correlate well with the time series AUC (correlation coefficient 0.95 for P2APP and 0.90 for P3APP). We compute the proportion of complete cases by taking the number of feature vector rows with complete information over the total number of rows. As is apparent from Fig. 3-10, the proportions have been increasing over the years for both datasets. This is likely due to better data reporting practices by drug developers, a possible consequence of FDAAA.

Next, we examine the 2011–2015 window. Fig. 3-11 and Tables 3.13 and 3.14 summarize the distributions of prediction scores for the P2APP and P3APP datasets. We observe very similar patterns to the static pipeline predictions above. The histograms, average scores, and binning of samples indicate that pairs that fail tend to have lower prediction scores than those that advance. This shows that our classifiers are indeed able to differentiate successful candidates.

Table 3.15 summarizes the top ten most informative variables in the 5NN-RF classifiers over the eight rolling windows. We find them to be largely consistent with those observed in the static case: the trial outcome and trial status are significantly associated with success; trial characteristics (such as accrual, duration, and number of identified sites), sponsor track record, and drug medium appear frequently in both scenarios.

As in the static case, we also train indication-group specific classifiers using rolling windows. Tables 3.16 and 3.17 summarize the results for selected indication groups in P2APP and P3APP, respectively (see Appendix B.6 for results of all other indication groups). Indication groups with small sample sizes tend to produce poor and unstable specialized classifiers (e.g., the musculoskeletal indication group in P2APP). This is expected because models trained on small training sets are more susceptible to

overfitting, especially when non-linear algorithms such as RF are used. In contrast, indication groups with larger sample sizes tend to give rise to rather good classifiers (e.g., anti-cancer in P2APP).

For comparison, we disaggregate performance by indication group. We find that these classifiers do not lose out to their specialized counterparts. In fact, our results show that the former tend to exhibit more stable performance across the seven windows, particularly on indication groups with small sample sizes. We hypothesize that classifiers trained on all data benefit from having access to larger datasets with greater diversity, and are thus able to make more informed predictions. This suggests that it may be more appropriate to rely on general classifiers, rather than specialized ones, for predictions over time where samples are spread out over multiple windows, since further filtering by indication group results in even smaller sample sizes.

Finally, we extract the top five P2APP pipeline drug candidates with the highest scores in each indication group as predicted by the 2011–2015 rolling-window model. Table 3.18 summarizes the results. We include only candidates that are still outstanding at the time of writing (neither discontinued nor approved). It is encouraging that many of these candidates (indicated in italics) have advanced beyond phase 2 testing since our analysis, indicating the predictive power of our models. We include an interactive version (illustration in Fig. 3-12) where readers can filter our pipeline predictions by indication group and probability of approval. Ultimately, all biopharma stakeholders can use such scores to rank and evaluate the potential risks and rewards of drug candidates.

3.5 Discussion

Drug development is an extremely costly process and the accurate evaluation of a candidate drug's likelihood of approval is critical to the efficient allocation of capital. Historical successes and failures contain valuable insights on the characteristics of high-potential candidates. Unfortunately, such data are often incomplete due to partial reporting by investigators and developers. Most analytic methods require

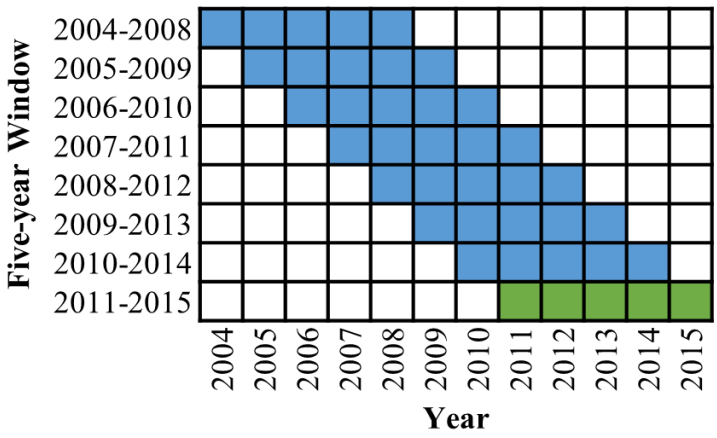


Figure 3-9: Time-series walk-forward analysis approach. The testing set in the last window (green) comprises drug-indication pairs in the pipeline at the time of snapshot of the databases.

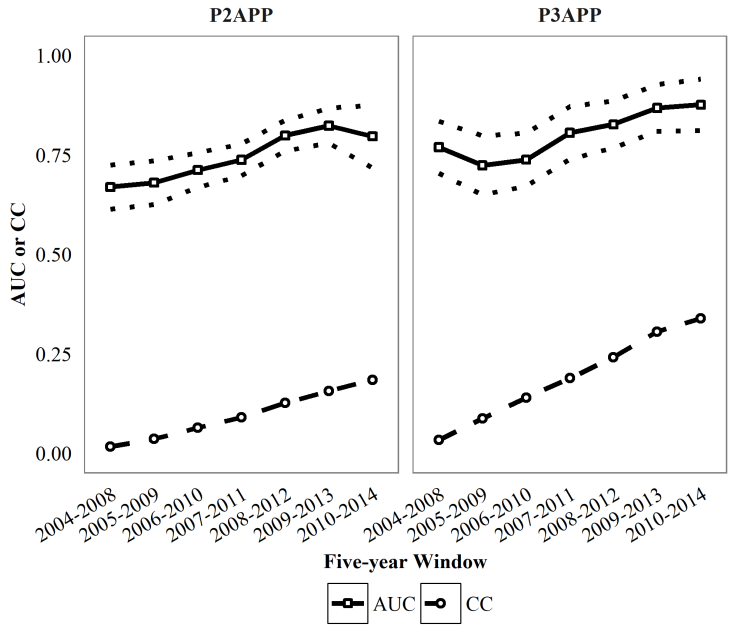


Figure 3-10: Time-series walk-forward analysis results for P2APP and P3APP using 5NN-RF. We use bootstrapping to determine the 95% CI for AUC (dotted lines). The dashed lines plot the corresponding proportions of complete cases in the training sets of each five-year window. *Abbreviations*: CC, proportion of complete cases.

Table 3.13: Distributions of prediction scores for all indication groups in aggregate. Advanced refers to progress to a higher phase from the original phase. Original phase for P2APP is phase 2; for P3APP is phase 3. *Abbreviations*: avg, average; sd, standard deviation; 5%, 5th percentile; 50%, median; 95%, 95th percentile; n, sample size.

	Prediction Scores					
	n	**Avg**	**Sd**	**5%**	**50%**	**95%**
			P2APP			
Aggregate	1,190	0.158	0.080	0.036	0.173	0.290
No change	712	0.148	0.080	0.035	0.158	0.275
Failed	195	0.143	0.079	0.034	0.149	0.255
Advanced	283	0.197	0.071	0.068	0.200	0.323
			P3APP			
Aggregate	218	0.431	0.211	0.113	0.476	0.689
No change	121	0.395	0.207	0.113	0.403	0.684
Failed	28	0.362	0.211	0.093	0.335	0.640
Advanced	69	0.521	0.193	0.149	0.631	0.707

Table 3.14: Distribution of prediction scores for all indication groups in aggregate. Proportion refers to the fraction of samples that advanced to a higher phase from the original phase. *Abbreviations*: n, sample size.

Scores	**n**	**Proportion**
	P2APP	
< 0.1	99	0.313
0.1-0.2	183	0.607
0.2-0.3	168	0.690
≥ 0.3	28	0.893
	P3APP	
< 0.2	17	0.412
0.2-0.4	17	0.706
0.4-0.6	17	0.647
≥ 0.6	46	0.848

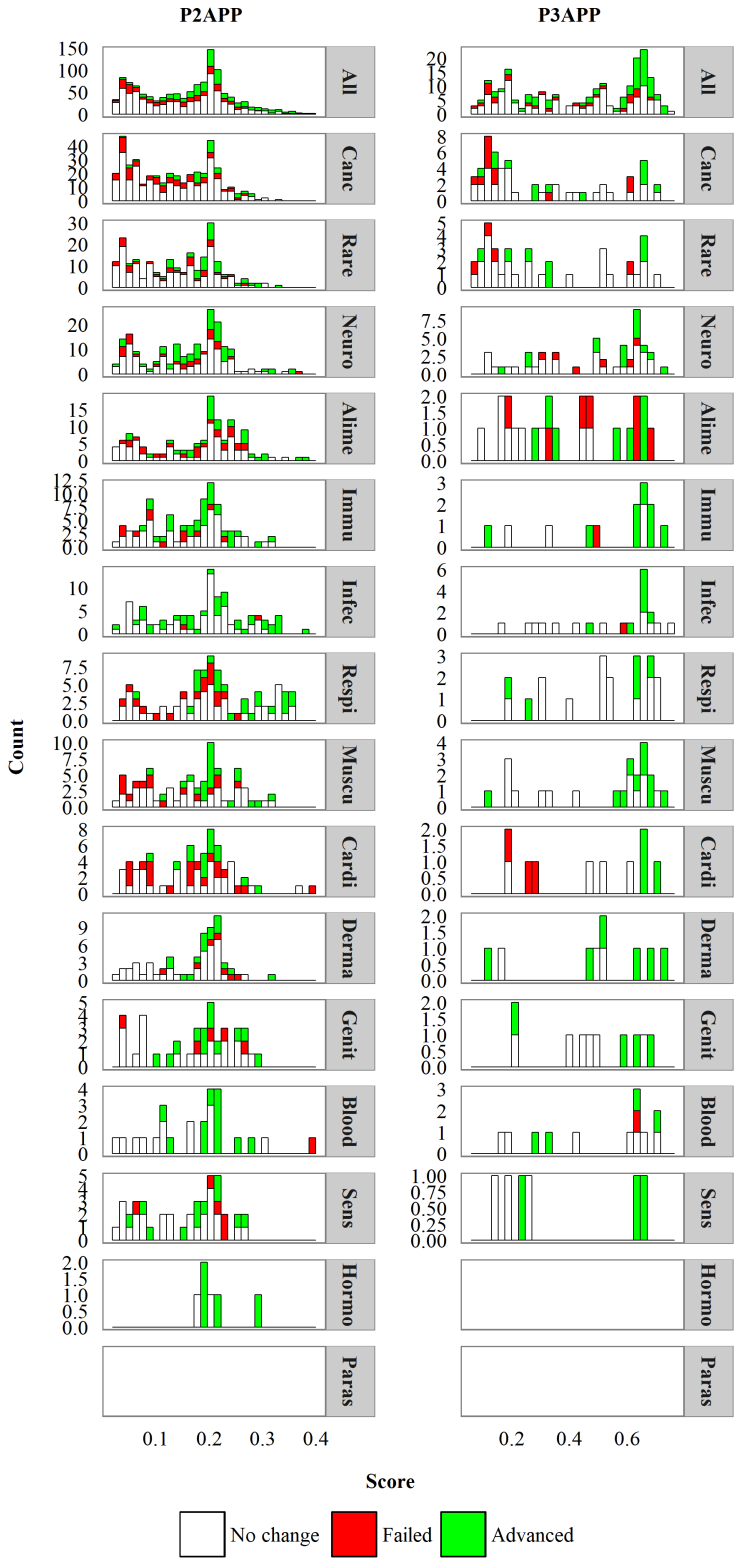


Figure 3-11: Distributions of prediction scores of the 2011–2015 window testing set for P2APP and P3APP. First row for all indication groups in aggregate. Subsequent rows for specific indication groups.

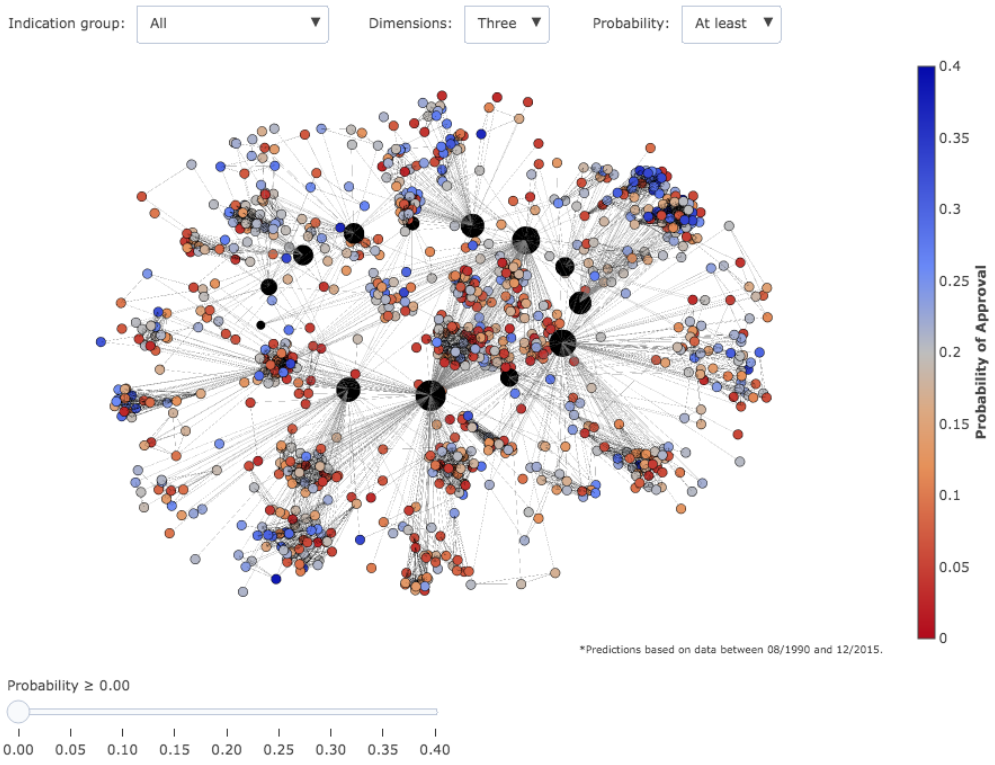


Figure 3-12: Network graph of P2APP pipeline drug candidates. See here for an interactive version. Black nodes correspond to indication groups. Colored nodes correspond to drug-indication pairs. Each drug-indication pair node is connected to its parent indication group and also other drug-indication pairs that have the same indication. They are colored according to their respective probability of approval as predicted by our model—blue for higher scores and red for lower scores. Hover over nodes for details of each drug-indication pair. Black indication group nodes are sized based on the number of connections.

Table 3.15: Top ten most important variables in 5NN-RF classifiers for P2APP and P3APP. Average and standard deviation taken across the eight rolling windows. *Abbreviations*: avg, average; sd, standard deviation.

	Importance	
	Avg	**Sd**
P2APP		
Trial outcome - completed, positive outcome or primary endpoint(s) met	0.203	0.083
Trial status	0.102	0.033
Prior approval of drug for another indication	0.077	0.061
Actual accrual	0.039	0.015
Target accrual	0.031	0.010
Duration	0.027	0.014
Sponsor track record - number of completed phase 3 trials	0.025	0.007
Medium - suspension	0.024	0.018
Sponsor type - academic	0.023	0.017
Medium - solution	0.021	0.019
P3APP		
Trial outcome - completed, positive outcome or primary endpoint(s) met	0.348	0.028
Trial status	0.125	0.020
Duration	0.053	0.017
Prior approval of drug for another indication	0.046	0.028
Trial outcome - completed, negative outcome or primary endpoint(s) not met	0.033	0.026
Target accrual	0.021	0.005
Trial outcome - terminated, lack of efficacy	0.020	0.013
Actual accrual	0.019	0.004
Therapeutic area - oncology	0.017	0.013
Number of identified sites	0.012	0.002

complete data, however, and prior studies on estimating approval rates and predicting approvals are typically based on a small number of examples that have complete information for just a few features.

In this study, we extract two datasets, P2APP and P3APP, from Informa® databases and apply 5NN statistical imputation to make efficient use of all available data. We use machine-learning techniques to train and validate our RF predictive models and achieve promising levels of predictive power for both datasets. When applied to pipeline drugs, we find that candidates with higher scores are indeed more likely to advance to higher clinical phases, indicating that our 5NN-RF classifiers are able to discriminate between high- and low-potential candidates.

A time-series analysis of the datasets shows generally increasing trends in performance over five-year rolling windows from 2004 to 2014. We find that the classifiers' performance correlates well with the proportions of complete cases in the training sets: as completeness increases, the classifier learns better and achieves higher AUCs.

Table 3.16: Out-of-sample and out-of-time performance for P2APP. Comparison of the general and indication-group specific classifiers for selected indication groups. We use bootstrapping to determine the 95% CI for AUC.

	General			Indication-Group-Specific		
	Training Set	**Testing Set**	**Testing Set AUC (95% CI)**	**Training Set**	**Testing Set**	**Testing Set AUC (95% CI)**
			All			
2004–2008	1,361	551	0.669 (0.614, 0.725)			
2005–2009	1,562	591	0.680 (0.625, 0.735)			
2006–2010	1,764	636	0.712 (0.668, 0.755)			
2007–2011	1,969	598	0.738 (0.698, 0.777)			
2008–2012	2,082	597	0.799 (0.760, 0.837)			
2009–2013	2,212	517	0.823 (0.779, 0.867)			
2010–2014	2,289	380	0.797 (0.718, 0.876)			
			Anti-cancer			
2004–2008	1,361	137	0.665 (0.528, 0.803)	456	137	0.683 (0.533, 0.833)
2005–2009	1,562	163	0.739 (0.618, 0.861)	494	163	0.635 (0.512, 0.758)
2006–2010	1,764	188	0.774 (0.702, 0.846)	546	188	0.726 (0.635, 0.816)
2007–2011	1,969	193	0.830 (0.773, 0.887)	618	193	0.746 (0.661, 0.831)
2008–2012	2,082	198	0.805 (0.717, 0.894)	682	198	0.760 (0.665, 0.855)
2009–2013	2,212	177	0.852 (0.783, 0.922)	736	177	0.786 (0.696, 0.876)
2010–2014	2,289	173	0.815 (0.691, 0.938)	791	173	0.803 (0.666, 0.940)
			Musculoskeletal			
2004–2008	1,361	35	0.765 (0.597, 0.933)	96	35	0.704 (0.512, 0.896)
2005–2009	1,562	38	0.716 (0.489, 0.944)	109	38	0.674 (0.472, 0.876)
2006–2010	1,764	35	0.634 (0.439, 0.830)	111	35	0.509 (0.276, 0.742)
2007–2011	1,969	37	0.737 (0.571, 0.903)	119	37	0.677 (0.493, 0.860)
2008–2012	2,082	36	0.884 (0.773, 0.995)	127	36	0.683 (0.462, 0.904)
2009–2013	2,212	26	0.792 (0.573, 1.000)	133	26	0.667 (0.429, 0.904)
2010–2014	2,289	19	0.882 (0.724, 1.000)	128	19	0.882 (0.706, 1.000)

Table 3.17: Out-of-sample and out-of-time performance for P3APP. Comparison of the general and indication-group specific classifiers for selected indication groups. We use bootstrapping to determine the 95% CI for AUC.

	General			Indication-Group-Specific		
	Training Set	**Testing Set**	**Testing Set AUC (95% CI)**	**Training Set**	**Testing Set**	**Testing Set AUC (95% CI)**
			All			
2004–2008	472	196	0.769 (0.704, 0.834)			
2005–2009	559	177	0.724 (0.650, 0.798)			
2006–2010	604	211	0.738 (0.671, 0.805)			
2007–2011	664	174	0.806 (0.740, 0.871)			
2008–2012	677	197	0.827 (0.768, 0.886)			
2009–2013	740	153	0.868 (0.809, 0.927)			
2010–2014	734	110	0.876 (0.811, 0.941)			
			Anti-cancer			
2004–2008	472	34	0.773 (0.618, 0.928)	95	34	0.684 (0.495, 0.874)
2005–2009	559	28	0.740 (0.543, 0.936)	107	28	0.568 (0.345, 0.791)
2006–2010	604	50	0.754 (0.599, 0.910)	110	50	0.630 (0.452, 0.809)
2007–2011	664	24	0.587 (0.333, 0.842)	132	24	0.392 (0.132, 0.651)
2008–2012	677	40	0.793 (0.549, 1.000)	134	40	0.668 (0.457, 0.879)
2009–2013	740	29	0.800 (0.480, 1.000)	151	29	0.775 (0.528, 1.000)
2010–2014	734	26	0.943 (0.842, 1.000)	153	26	0.852 (0.558 ,1.000)
			Rare Diseases			
2004–2008	472	22	0.711 (0.465, 0.957)	54	22	0.620 (0.364, 0.876)
2005–2009	559	23	0.735 (0.517, 0.952)	60	23	0.606 (0.360, 0.852)
2006–2010	604	24	0.888 (0.747, 1.000)	66	24	0.825 (0.645, 1.000)
2007–2011	664	22	0.838 (0.652, 1.000)	72	22	0.735 (0.520, 0.950)
2008–2012	677	34	0.893 (0.780, 1.000)	76	34	0.700 (0.523, 0.877)
2009–2013	740	28	0.962 (0.899, 1.000)	94	28	0.932 (0.840, 1.000)
2010–2014	734	18	0.908 (0.766, 1.000)	109	18	0.985 (0.942, 1.000)

Table 3.18: Top five P2APP pipeline drug candidates with the highest scores in each indication group as predicted by our model. We include only candidates that are still outstanding at the time of writing (neither discontinued nor approved). Drug-indication pairs in *italics* are those that have advanced beyond phase 2 testing since our analysis.

Drug	Indication	Score	Drug	Indication	Score
Anti-cancer			**Musculoskeletal**		
ontecizumab	Cancer, colorectal	0.34	*tofacitinib*	*Arthritis, psoriatic*	*0.31*
calmangafodipir	Radio/chemotherapy-induced injury, bone marrow, neutropenia	0.31	ixekizumab	Arthritis, rheumatoid	0.31
tivantinib	Cancer, sarcoma, soft tissue	0.30	anti-BLyS/APRIL antibody fusion protein	Arthritis, rheumatoid	0.31
pidilizumab	Cancer, colorectal	0.29	*sirukumab*	*Arthritis, rheumatoid*	*0.29*
NK-012	Cancer, colorectal	0.28	*romosozumab*	*Osteoporosis*	*0.28*
Rare Diseases			**Cardiovascular**		
surotomycin	*Infection, Clostridium difficile*	*0.34*	K-134	Peripheral vascular disease	0.37
tivantinib	Cancer, sarcoma, soft tissue	0.30	*nitric oxide, inhaled*	*Hypertension, pulmonary*	*0.29*
VP-20621	Infection, Clostridium difficile prophylaxis	0.30	TY-51924	Infarction, myocardial	0.28
KHK-7580	*Secondary hyperparathyroidism*	*0.29*	*s-amlodipine + telmisartan*	*Hypertension, unspecified*	*0.27*
nitric oxide, inhaled	*Hypertension, pulmonary*	*0.29*	tirasemtiv	Peripheral vascular disease	0.24
Alimentary			**Genitourinary**		
dasotraline	*Attention deficit hyperactivity disorder*	*0.35*	*tofacitinib*	*Arthritis, psoriatic*	*0.31*
idalopirdine	*Alzheimer's disease*	*0.35*	dimethyl fumarate	Psoriasis	0.27
GRC-17536	Neuropathy, diabetic	0.34	*pefcalcitol*	*Psoriasis*	*0.24*
caprylic triglyceride	*Alzheimer's disease*	*0.32*	*Benvitimod*	*Psoriasis*	*0.22*
levodopa	*Parkinson's disease*	*0.31*	calcipotriol monohydrate + betamethasone dipropionate	Psoriasis	0.22
Neurological			**Dermatological**		
ibodutant	*Irritable bowel syndrome, diarrhoea-predominant*	*0.37*	*etonogestrel + estradiol (vaginal ring), next generation*	*Contraceptive, female*	*0.30*
GRC-17536	Neuropathy, diabetic	0.34	drospirenone + estradiol	Contraceptive, female	0.28
mesalazine + N-acetylcysteine	Colitis, ulcerative	0.31	*finerenone*	*Nephropathy, diabetic*	*0.27*
apabetalone (tablet)	*Diabetes, Type 2*	*0.31*	afacifenacin fumarate	Overactive bladder	0.26
phosphatidylcholine	*Colitis, ulcerative*	*0.31*	GKT-137831	Nephropathy, diabetic	0.26

Table 3.18 (continued): Top five P2APP pipeline drug candidates with the highest scores in each indication group as predicted by our model. We include only candidates that are still outstanding at the time of writing (neither discontinued nor approved). Drug-indication pairs in *italics* are those that have advanced beyond phase 2 testing since our analysis.

Drug	Indication	Score	Drug	Indication	Score
Immunological			**Blood and Clotting**		
tofacitinib	*Arthritis, psoriatic*	*0.31*	calmangafodipir	Radio/chemotherapy-induced injury, bone marrow, neutropenia	0.31
ixekizumab	Arthritis, rheumatoid	0.31	*balugrastim*	*Radio/chemotherapy-induced injury, bone marrow, neutropenia*	*0.27*
anti-BLyS/APRIL antibody fusion protein	Arthritis, rheumatoid	0.31	*eflapegrastim*	*Radio/chemotherapy-induced injury, bone marrow, neutropenia*	*0.25*
sirukumab	*Arthritis, rheumatoid*	*0.29*	*pegfilgrastim*	*Radio/chemotherapy-induced injury, bone marrow, neutropenia*	*0.22*
dimethyl fumarate	Psoriasis	0.27	lexaptepid pegol	Radio/chemotherapy-induced anaemia	0.20
Anti-infective			**Sensory**		
delafloxacin	*Infection, skin and skin structure, acute bacterial*	*0.39*	*AR-13324 + latanoprost*	*Glaucoma*	*0.27*
surotomycin	*Infection, Clostridium difficile*	*0.34*	S-646240	Macular degeneration, age-related, wet	0.27
delafloxacin	*Infection, pneumonia, community-acquired*	*0.33*	*netarsudil*	*Glaucoma*	*0.26*
plazomicin	*Infection, urinary tract, complicated*	*0.33*	fenofibrate, micronized-2	Oedema, macular, diabetic	0.25
Ypeginterferon alpha-2b	*Infection, hepatitis-C virus*	*0.33*	LX-7101	Glaucoma	0.21
Respiratory			**Hormonal**		
fluticasone + salmeterol	*Asthma*	*0.36*	KHK-7580	Secondary hyperparathyroidism	0.29
fluticasone furoate + umeclidinium + vilanterol	*Chronic obstructive pulmonary disease*	*0.36*	*somatropin prodrug, pegylated*	*Growth hormone deficiency*	*0.21*
fluticasone furoate + umeclidinium	Chronic obstructive pulmonary disease	0.36	2MD	Secondary hyperparathyroidism	0.21
beclometasone + formoterol	Chronic obstructive pulmonary disease	0.35	*velcalcetide*	*Secondary hyperparathyroidism*	*0.19*
fluticasone propionate DPI	*Asthma*	*0.35*	tesamorelin acetate	Growth hormone deficiency	0.18

This highlights the importance of data quality in building more accurate predictive algorithms for drug development.

Finally, we compute feature importance in the predictive models and find that trial outcomes, trial status, trial accrual rate, duration, prior approval for another indication, and sponsor track record are the most critical features for predicting success. Because the 5NN-RF classifiers are non-linear, there is no simple interpretation of the incremental contribution of each predictor to the forecast. However, the intuition behind some of these factors is clear: drug-indication pairs with trials that achieve positive outcomes certainly have a better chance of approval; candidates sponsored by companies with strong track records and greater expertise in drug development should have higher likelihood of success; and approved drugs may have higher chances of approval for a second related indication. Many of these factors contain useful signals about drug development outcomes but have not been considered in prior studies.

These results are promising and raise the possibility of even more powerful drug development prediction models with access to better quality data. This can be driven by programs such as Project Data Sphere [84] and Vivli [85] that promote and facilitate public sharing of patient-level clinical trial data. Ultimately, such predictive analytics can be used to make more informed data-driven decisions in the risk assessment and portfolio management of investigational drugs at all clinical stages.

Chapter 4

Cost/Benefit Analysis of Vaccine Trial Designs for COVID-19

The world is facing unprecedented challenges from the COVID-19 pandemic. It is clear that the development of a vaccine is critical to stopping the epidemic and the socio-economic crisis. Given the dire situation, human challenge clinical trials (HCTs) have been proposed as a way to accelerate the vaccine development process, which typically takes more than a decade to complete. While moral concerns have been raised, bioethicists generally agree that an HCT may be ethically permissible if one can demonstrate the explicit societal value of an HCT. However, to the best of our knowledge, there has not been any quantitative analysis of the societal value of a COVID-19 HCT versus non-challenge trials in literature, thus making it difficult to justify the use of a challenge study at this time. In this chapter, we propose a simulation framework for quantitatively assessing the costs and benefits trade-offs—as measured by the expected number of infections and deaths that can be avoided—of four vaccine efficacy clinical trial designs for COVID-19, including an HCT. Using epidemiological models calibrated to the current pandemic, we simulate the time course of each trial design for a total of 756 unique combinations of parameters—such as different vaccine efficacies, epidemiological scenarios, vaccination schedules after licensure, approval requirements, and set-up times for HCTs—to determine which design is most effective for a given scenario. We find that a human challenge trial

provides maximal net benefits—averting up to 1.1M infections and 8,000 deaths in the U.S. compared to the next best non-challenge clinical trial design—if initiated early in an epidemic or if the rate of infection is relatively low. In most of the other cases, we find that an adaptive trial provides greater net benefits. This framework will allow stakeholders to make more informed practical and ethical decisions regarding accelerating COVID-19 vaccine development in the ongoing pandemic.

4.1 Introduction

The COVID-19 pandemic has caused the deaths of hundreds of thousands. Its economic fallout has upended the lives of billions, and caused trillions of dollars in financial losses. Life may not return to normal until a vaccine is found [86]. Despite the many candidates undergoing testing, an approved vaccine is not expected until 2021, even with substantially compressed development timelines [87], smooth proceeding of clinical trials, and not accounting for possible failures [18]. It is possible—though considered highly unlikely at the present time—that, like many non-influenza epidemics, the crisis may be over before a successful vaccine is developed [88].

Unlike typical therapeutics that are administered to sick patients, vaccines are intended for the healthy. Therefore, confirming the safety and effectiveness of a vaccine is of critical importance [89]. The two primary methods for demonstrating vaccine safety and efficacy are through either a vaccine efficacy randomized clinical trial (RCT) or a vaccine immunogenicity RCT. In the former, large numbers of healthy volunteers are randomly selected to receive either a vaccine candidate or a control and then monitored for a period of time. At the end of the surveillance period, the difference in the proportion of infections between the treatment and control arms is computed to demonstrate the ability of the vaccine to prevent infection or disease. A large number of participants and a long time is needed to obtain statistically significant results because only a small proportion of participants will eventually encounter the disease since many will take precautions to avoid exposure. To put things into perspective, a phase 3 vaccine efficacy RCT typically takes five to ten

years to complete [90].

In a vaccine immunogenicity RCT, the primary endpoint is an immunity measurement or a surrogate marker that is known to correlate with protection against infection or a disease. This type of trial involves a smaller number of volunteers and requires a shorter follow-up period, and as a result, is quicker and less costly [91]. Given that SARS-CoV-2 is a novel pathogen for which we do not yet know how to determine whether a subject is protected, vaccine efficacy must be confirmed using the longer and more costly vaccine efficacy RCT. While there exists the possibility of an expedited (conditional) licensure based on immunogenicity results with post-approval commitments, we find it unlikely to occur given the latest information. The U.S. Food and Drug Administration (FDA) has also issued a guidance stating that "the goal of development programs should be to pursue traditional approval via direct evidence of vaccine efficacy [92]."

Given the dire situation posed by COVID-19, human challenge clinical trials (HCTs) have been proposed as an alternative to expedite the vaccine development process. In HCTs, participants are randomized into either the vaccine or control arm, and then deliberately exposed to a virus under controlled conditions to study safety and efficacy. In comparison to traditional vaccine efficacy field trials, HCTs typically require a much smaller sample size and shorter duration as investigators do not need to wait for infections to occur naturally. Moreover, correlates of protection—that is, measurable signs that a person is immune from infection and/or the disease, e.g., minimum level of antibodies for immunity—can be more easily established in HCTs [91]. This can help to accelerate future COVID-19 vaccine development because correlates of protection can be used as surrogate endpoints for vaccine efficacy in immunogenicity trials, which can be completed much faster.

While concerns have been raised regarding the ethics and morality of HCTs, it turns out that ethically permissible HCTs have been successfully conducted in the past for multiple infectious diseases such influenza [93], cholera [94], malaria [95], typhoid fever [96], and dengue fever [97]. In fact, there are already multiple studies in literature that examine the ethical considerations of HCTs for COVID-19 vaccine

development. One fundamental premise of many of these discussions is the high social value of challenge trials versus traditional pathways [98]. While the estimation of benefits is clearly a key criterion for justifying COVID-19 HCTs, we do not find any studies in literature that goes beyond a qualitative discussion to quantify the potential value of HCTs.

In this chapter, we seek to bridge this gap by proposing a systematic framework for quantitatively assessing the costs and benefits of four different clinical trial designs for COVID-19 vaccine development, including a traditional vaccine efficacy RCT, a vaccine efficacy RCT with an optimized fixed-duration surveillance period that maximizes the benefits of the trial (ORCT), an adaptive vaccine efficacy RCT (ARCT), and an HCT.

4.2 Simulation Framework

Our simulation framework is illustrated in Fig. 4-1. In our analysis, we quantify cost/benefit in terms of the number of infections prevented and deaths averted. Although our framework applies broadly to any vaccine candidate for any infectious disease, we calibrate our simulations to the current pandemic.

We first estimate baseline epidemiological models for COVID-19 in U.S. and consider several possible scenarios regarding the evolution of the epidemic after the reopening and during the clinical trial. Using the attack rates from the estimated epidemic model—the proportion of a susceptible population infected with a disease, we simulate the outcomes of different clinical trial designs, including the date of licensure and the probability of approval. Conditioned on the approval of the vaccine candidate, we make assumptions on the vaccination schedule and simulate the new path of the epidemic in order to compute the net number of infections and deaths prevented versus the baseline case where no vaccine is ever approved. We compute the expected net value of each clinical trial design using Monte Carlo simulations.

We review the assumptions of four vaccine trial designs in Section 4.3, review the statistical methods for efficacy analysis in Section 4.4, and present the epidemiological

model used in our forecasts in Section 4.5. We describe our cost/benefit computations in Section 4.6 and report our simulation results in Section 4.7. Finally, we discuss our findings and some broader issues of COVID-19 clinical trials in Section 4.8 and conclude in Section 4.9.

4.3 Vaccine Efficacy Trial Designs

4.3.1 Traditional Randomized Clinical Trial

First, we consider a traditional double-blind vaccine efficacy trial design. We assume that a closed cohort of 30,000 infection-free but at-risk healthy U.S. adults aged between 18 and 50 years will be enrolled for the study, similar to the phase 3 studies planned or underway for the COVID-19 vaccines developed by Moderna [99], AstraZeneca [100], Pfizer/BioNTech [101], and others. The participants will be randomized equally between the treatment and control arms, receiving either the vaccine candidate or an active control vaccine (e.g., vaccine against meningococcal bacteria), respectively. The use of an active vaccine (e.g., vaccine against meningococcal bacteria) as control provides some benefit to the participants, making it more ethical. It also serves to ensure that the participants are unable to tell whether they received the COVID-19 vaccine based on side effects such as soreness at the injection site, reducing the possibility of behavioral changes that can bias the results of the study.

Unlike clinical trials for cancer therapeutics where patient accrual can be a challenge due to the small pool of afflicted patients and strict inclusion/exclusion criteria, subject enrollment for vaccine efficacy studies are often accelerated because there is a large pool of healthy adult volunteers to recruit from. Therefore, we assume an accrual rate of 250 patients per day in our simulations.

Similar to the design of study protocols adopted for phase 3 clinical trials of current leading SARS-CoV-2 vaccine candidates, we assume a hypothetical COVID-19 vaccine candidate that will be administered to subjects in two doses, 28 days apart, i.e., the prime-boost regimen [102, 103]. Furthermore, we assume that it takes approximately

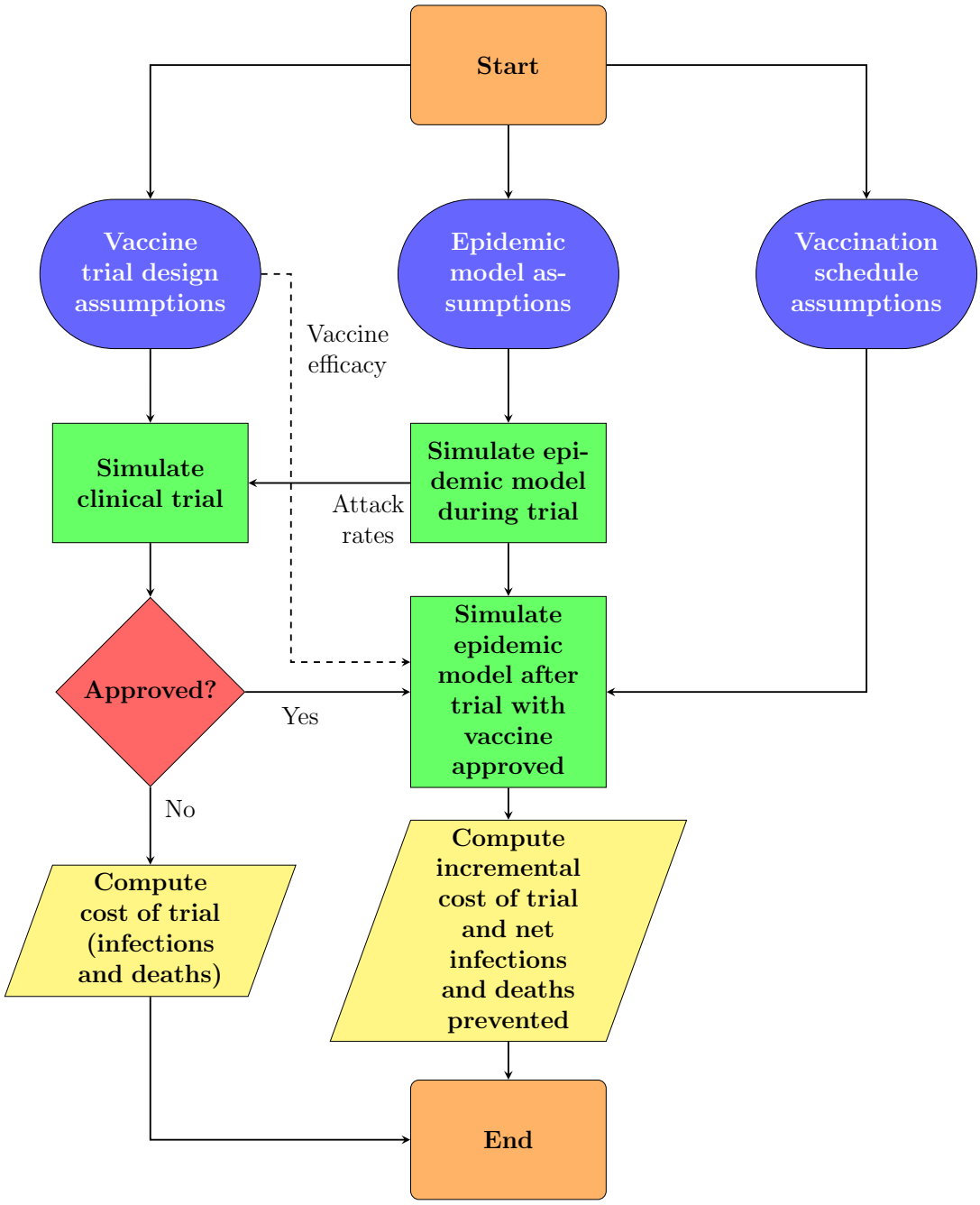


Figure 4-1: Simulation framework. For each Monte Carlo simulation path, we simulate patient-level infections data based on input trial design assumptions and attack rates from the population epidemiological model (for an RCT, ORCT, and ARCT). At the end of the trial (or, at each interim analysis for an ARCT), we determine if the vaccine candidate is approved under superiority or superiority-by-margin testing. Finally, we compute the expected net value of the trial design over 100,000 simulation paths.

28 days after the booster dose for antibodies to develop (i.e., seroconversion) before surveillance can begin.

We consider efficacy in the prevention of infection by SARS-CoV-2 as the primary endpoint in our study. To draw meaningful conclusions from the trial results, volunteers must be monitored long enough for a sufficient number of infections to occur. Here, we assume a fixed post-vaccination surveillance period of 180 days for all subjects in the cohort, after which a single safety and primary efficacy analysis will be performed to determine licensure (see Section 4.4.1).

Finally, we assume an interval of 120 days after surveillance for the preparation of a biologics license application (BLA) submission to the FDA, including an analysis and publication of safety, immunogenicity, and efficacy results; collection of chemistry, manufacturing, and controls (CMC) data; the writing of a clinical study report; and subsequent review by the FDA. Under these assumptions, we estimate the time to licensure of our hypothetical candidate under a traditional RCT to be approximately 476 days. This is the baseline value against which we will compare the other three trial designs.

4.3.2 Optimized Randomized Clinical Trial

Depending on the transmission rate of COVID-19 during the trial and the assumed efficacy of the hypothetical candidate, a shorter surveillance period might be sufficient to observe significant results. In general, the higher the transmission rate, the shorter the surveillance period required to observe a statistically significant difference in infection risk between the treatment arm and the control arm (or the lack of thereof) at the same level of significance and power, assuming a constant sample size and vaccine efficacy.

We also note that there is a trade-off between time and power here: A shorter surveillance period will, *ceteris paribus*, reduce the power of the RCT. However, it will also reduce the time to licensure of the vaccine (if approved), which can potentially prevent more infections and save more lives. Conversely, a longer surveillance period will increase the power of the RCT but prolong the time it takes for the vaccine to

be approved.

Therefore, we consider an optimized version of the traditional vaccine efficacy RCT design in which the surveillance period is optimized between 30 to 180 days based on different epidemiological scenarios and vaccine efficacies to maximize the expected number of incremental infections and deaths prevented. Apart from the surveillance period, we assume that the ORCT is identical to the RCT in all other aspects.

4.3.3 Adaptive Randomized Clinical Trial

An adaptive version of the traditional vaccine efficacy RCT design is based on group sequential methods [104]. Instead of a fixed study duration with a single final analysis at the end, we allow for early stopping for efficacy via periodic interim analyses of accumulating trial data (see Section 4.4.3). While this reduces the expected duration of the trial, adaptive trials typically require more complex study protocols which can be operationally challenging to implement for test sites unfamiliar with this framework.

In our simulations, we assume a maximum of six interim analyses spaced 30 days apart, with the first analysis performed when the first 10,000 subjects have been monitored for at least 30 days. While we have assumed interim analyses at periodic calendar time points here, we note that most vaccine efficacy trials are event based, e.g., performing interim analyses when pre-specified numbers of events occur. In addition, we have adopted Pocock's test for sequential testing (see Section 4.4.3), but we note that some companies are using variants of the O'Brian-Fleming test [105], which have stricter requirements for early stopping, and therefore may lead to longer studies [104].

4.3.4 Human Challenge Trial

Unlike traditional vaccine efficacy field trials which require large sample sizes to observe significant results, we assume that the HCT requires only 250 volunteers, randomized 4:1 between the treatment and control arms. Furthermore, to minimize the

risk to participants, we assume that this study will recruit only young and healthy adults aged between 18 and 25 years without any underlying chronic conditions because this group of individuals has the lowest risk of mortality and complications after recovering from the infection [106, 107, 108].

It is clear that extensive preparations are required to set up an HCT: selecting, developing, and testing an appropriate challenge virus strain among multiple lineages of SARS-CoV-2; manufacturing a batch of the selected challenge strain under good manufacturing practices (GMP); and identifying the dose level required to achieve satisfactory attack risk of non-severe clinical illness [108] (see Appendix C.7). From discussions with challenge trial experts, there seems to be a lack of consensus on the appropriate set-up time for HCTs. We reflect this uncertainty in our simulations by incorporating a lag time for HCTs ("set-up time") that ranges between 30 to 120 days.

In the challenge study, volunteers are deliberately exposed to the SARS-CoV-2 virus, reducing post-vaccination monitoring times because investigators do not need to wait for infections to occur naturally as with non-challenge RCTs. Therefore, we assume a surveillance period of only 14 days (the incubation period for COVID-19 [109, 110, 111]) for the challenge study. Moreover, the attack rate in the control arm will be independent of the population epidemiological model since the study will be conducted in isolated facilities. In our simulations, we assume that 90% of the subjects in the control arm will be infected after the challenge. We do not assume a 100% attack rate since the challenge strain used is likely weakened to reduce risk to volunteers, and some individuals might have innately stronger immune systems that can counteract the virus.

We note that the FDA is unlikely to approve an experimental vaccine tested in only 200 subjects (versus thousands in non-challenge RCTs), hence we assume that a large-scale safety study will be performed immediately after the conclusion of the challenge study—conditional on positive efficacy results—to evaluate the safety of the hypothetical vaccine candidate in a broader population. Assuming a single-arm study with 5,000 subjects followed for 30 days, we expect the process to be completed

Table 4.1: Trial design assumptions common across RCT, ORCT, ARCT, and HCT.

Parameter	**Value**
Cohort	Closed and fixed
Accrual rate (patients/day)	250
Control arm	Vaccine for meningococcal bacteria
Treatment arm	Vaccine candidate for COVID-19
Vaccination schedule	Two doses administered 28 days apart
Vaccine efficacy (%)	30–90
Time for immune response (days)	28
Endpoint	Infection by SARS-CoV-2
Time for safety data collection data analysis, and FDA review (days)	120
Type I error (%)	5

in 106 days. To accelerate licensure, we assume that the collection of safety data will be performed in parallel with BLA submission and FDA review. Since the latter is assumed to take 120 days, the additional safety study does not actually add to the time to licensure of the vaccine candidate. It does, however, add to the financial costs of the HCT (see Appendix C.5).

Apart from the sample size, randomization ratio, set-up time, surveillance period, and safety data requirement, we assume that the HCT is identical to the RCT in all other respects. See Tables 4.1 and 4.2 for a summary of our assumptions.

We anticipate similar post-marketing commitments for both the HCT and the non-challenge RCTs, in terms of the collection of additional safety and effectiveness data, and supplementary studies to support the effectiveness of the vaccine in populations not included in the initial efficacy study, e.g., infants. However, we do not model them here because they do not affect our cost/benefit computations.

4.4 Efficacy Analysis

4.4.1 Fixed-Duration Clinical Trial

The protective effect of a vaccine—that is, vaccine efficacy—is defined as [91]:

$$\varepsilon \quad = \quad 1 - \frac{p_1}{p_0} \quad = \quad 1 - \frac{c_1/n_1}{c_0/n_0} \tag{4.1}$$

where ε refers to the vaccine efficacy, p_1 and p_0 are the attack rates observed in

Table 4.2: Trial design assumptions specific to RCT, ORCT, ARCT, and HCT.

Parameter	RCT	ORCT	ARCT	HCT
Set-up time (days)				30–120
Sample size	30,000	30,000	30,000	250
Inclusion criteria	Healthy adults aged 18–50 years	Healthy adults aged 18–50 years	Healthy adults aged 18–50 years	Healthy adults aged 18–25 years
Randomization ratio (treatment:control)	1:1	1:1	1:1	4:1
Time for enrollment (days)	120	120	40–120	1
Surveillance period (days)	Fixed and constant for all subjects; 180	Fixed and constant for all subjects; 30–180	Calendar time interval	Fixed and constant for all subjects; 14
Attack rate (%)	Depends on epidemiological model	Depends on epidemiological model and surveillance period	Depends on epidemiological model and surveillance period	90
Efficacy analysis	Single analysis at end of study	Single analysis at end of study	Up to 6 interim analyses spaced 30 days apart	Single analysis at end of study
Additional safety study				Single-arm with 5,000 subjects
Estimated time to licensure (days)	476	326–476	246–396	221–311

the treatment arm and the control arm, respectively, n_1 and n_0 refer to the sample sizes of the treatment arm and the control arm, respectively, and c_1 and c_0 refer to the number of infections observed in the treatment arm and the control arm, respectively. The attack rate is defined as the fraction of a cohort at risk that becomes infected during the surveillance period. There are conflicting views on the possibility of human reinfections [112, 113]; for simplicity, we rule out recurrent infections in our simulations.

Superiority Testing

First, we consider superiority testing to determine the licensure of a vaccine candidate at the end of a clinical study, e.g., RCT, ORCT, or HCT. The aim is to demonstrate that the efficacy of the candidate in the prevention of infections is greater than zero. Such a criteria might be appropriate for emergency use authorization during a pandemic where no alternative treatments are available. For this, we consider the following null and alternative hypotheses:

$$H_0 : p_1 - p_0 \geq 0 \qquad , \qquad H_1 : p_1 - p_0 < 0 \tag{4.2}$$

The test statistic under the null hypothesis is given by:

$$
\begin{aligned}
z &= \frac{|p_1 - p_0| - a}{\sqrt{2\bar{p}\bar{q}a}} \\
a &= \frac{r+1}{2rn_0} \quad , \quad r = \frac{n_1}{n_0} \\
\bar{p} &= \frac{c_1 + c_0}{n_0(r+1)} = \frac{rp_1 + p_0}{r+1} \quad , \quad \bar{q} = 1 - \bar{p}
\end{aligned}
\tag{4.3}
$$

where z is the test statistic. For large samples, z is approximately the standard Normal distribution.

The power of a vaccine efficacy study under superiority testing is given by [114, 115]:

$$
\begin{aligned}
z_\beta &= \frac{|P_1 - P_0|\sqrt{rn_0 - (r+1)/|P_1 - P_0|} - z_{\alpha/2}\sqrt{(r+1)\bar{P}\bar{Q}}}{\sqrt{P_1Q_1 + rP_0Q_0}} \\
\bar{P} &= \frac{rP_1 + P_0}{r+1} \quad , \quad \bar{Q} = 1 - \bar{P} \\
P_1 &= (1-\epsilon)P_0 \quad , \quad Q_i = 1 - P_i, \quad i \in \{0, 1\}
\end{aligned}
\tag{4.4}
$$

where α is the level of significance, β refers to the type II error under the alternative hypothesis, z_a is the $100(1-a)$ percentage points of the standard Normal distribution, P_1 and P_0 refer to the underlying attack rate in the treatment arm and the control arm, respectively, and ϵ refers to the true vaccine efficacy.

4.4.2 Superiority-by-Margin Testing

Next, we consider the case where superiority by margin (also known as super-superiority)—that is, a vaccine efficacy that is greater than some minimum threshold—must be demonstrated for full licensure:

$$
H_0 : \vartheta - \theta \geq 0 \quad , \quad H_1 : \vartheta - \theta < 0 \tag{4.5}
$$

where $\vartheta = p_1/p_0$, and θ is a specified minimum threshold larger than 0 and smaller than 1.

The test statistic under the null hypothesis is given by [114]:

$$z = \frac{|p_1 - \theta p_0|}{\sqrt{(\tilde{p}_1\tilde{q}_1 + r\theta^2\tilde{p}_0\tilde{q}_0)/rn_0}} \tag{4.6}$$
$$\tilde{q}_i = 1 - \tilde{p}_i, \qquad i \in \{0, 1\}$$

where z is the test statistic, and $\tilde{p}_1$ and $\tilde{p}_0$ are the large sample approximations of the constrained maximum likelihood estimate of P_1 and P_0, respectively, under the null hypothesis (see Appendix C.1 for closed-form solutions). For large samples, z is approximately the standard Normal distribution.

The power of a vaccine efficacy study under superiority-by-margin testing is given by:

$$z_\beta = \frac{(\theta P_0 - P_1)\sqrt{rn_0} - z_{\alpha/2}\sqrt{\tilde{p}_1\tilde{q}_1 + r\theta^2\tilde{p}_0\tilde{q}_0}}{\sqrt{P_1Q_1 + r\theta^2P_0Q_0}} \tag{4.7}$$

4.4.3 Adaptive Clinical Trial

We propose an adaptive vaccine efficacy RCT design (ARCT) based on group sequential methods. First, we consider an alternative definition of vaccine efficacy based on relative force of infection, as opposed to relative risk of infection in Eq. (4.1):

$$\varepsilon \approx 1 - \frac{\Lambda_1}{\Lambda_0} \tag{4.8}$$
$$\Lambda_i = \int_0^{t_s} \lambda_i(u)\,\mathrm{d}u, \qquad i \in \{0, 1\}$$

where λ_1 and λ_0 refer to the force of infection in the treatment arm and the control arm, respectively, and t_s refers to the duration of the surveillance period. The force of infection of an infectious disease is defined as the expected number of new cases of the disease per unit person-time at risk. When the risk of infection is small, e.g., smaller than 0.10, the risk of infection is approximately equal to the cumulative force of infection [91].

Next, we note that the force of infection and the hazard function in survival

analysis actually take the same functional form [91]. This suggests that infections can also be treated as time-to-event data, in addition to binary variables as in Eq. (4.1). By performing Cox regression on the time-to-infections data of a clinical trial, we can estimate the efficacy of the vaccine candidate from the hazard ratio of the treatment arm versus the control arm:

$$\varepsilon \approx 1 - \exp(\beta) \tag{4.9}$$
$$\lambda(t|z) = \lambda_{\text{baseline}}(t) \exp(\beta z)$$

where z refers to the treatment variable, i.e., whether the patient is vaccinated or not, $\lambda_{\text{baseline}}$ is the baseline hazard function, and β is the log hazard ratio. We note that the proportional hazards assumption is not unreasonable if we assume that the proportion of cases prevented by the vaccine is independent of the possibly non-homogeneous force of infection [91].

We consider the following null and alternative hypotheses based on the coefficient of the treatment variable in the Cox model:

$$H_0 : \beta - \beta_0 \geq 0 \qquad , \qquad H_1 : \beta - \beta_0 < 0 \tag{4.10}$$

where β_0 is 1 for superiority testing and smaller than 1 for superiority-by-margin testing.

The test statistic under the null hypothesis is given by:

$$z = \frac{|\hat{\beta} - \beta_0|}{\text{se}(\hat{\beta})} \tag{4.11}$$

where $\hat{\beta}$ is the maximum partial likelihood estimate of β and $\text{se}(\hat{\beta})$ is its standard error, and z is asymptotically Normal. This is also known the Wald test. It turns out this statistic satisfies the criteria for group sequential testing [104], allowing us to perform periodic interim analyses of accumulating trial data, rather than just a single final analysis at the end of a traditional vaccine efficacy RCT (see Fig. 4-2).

Under the group sequential testing framework, we estimate a new Cox model at

each interim calendar time point based on the infections data that has accrued up to that point, over the course of the study surveillance period. At the interim analyses, we decide whether to stop the study early by rejecting the null hypothesis, i.e., approving the vaccine candidate, or to continue on to the next analysis by monitoring the subjects for a longer period of time [104].

We adopt Pocock's test for sequential testing [116]. It involves repeated testing at successive interim analyses at some constant nominal significance level over the course of the study (see Algo. 4-1). The critical value is chosen to satisfy the maximum type I error requirement, e.g., 5%.

In our simulations, we consider a maximum of six interim analyses spaced 30 days apart, with the first analysis performed when the first 10,000 subjects enrolled have been monitored for at least 30 days. To keep the type I error at 5%, we consider a nominal significance level of 2.453 at each interim analyses [116].

For each of the epidemiological-model and population-vaccination schedule assumptions, we compute the expected net value of ARCT over 100,000 Monte Carlo simulation paths. For each path, we track the infections data of 30,000 patients for up to 180 days of surveillance. In addition, we estimate up to six Cox proportional hazards models, one at each interim analysis. The simulation process is computationally intensive despite parallelization, requiring approximately 8 hours to complete on the MIT Sloan "Engaging" high-performance computing cluster using over 400 processors.

While we have considered a simple adaptive design in this analysis, we note that our framework can be easily extended to other sequential boundaries such as the O'Brien & Fleming's test, to two-sided tests that allow for early stopping under the null hypothesis, i.e., early stopping for both futility and efficacy, and to flexible monitoring using the error spending approach, instead of using a constant nominal significance level for all interim analyses [104].

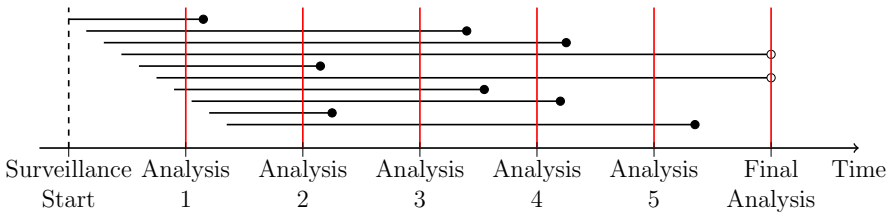


Figure 4-2: Infections as time-to-event data, measured from the start of surveillance. The horizontal lines represent the time to infection of ten subjects enrolled at different times. We monitor the subjects until an infection occurs or the end of study, whichever comes earlier. A solid circle at the right end denotes an infection, whereas a hollow circle indicates censoring. In the figure, we consider up to six analyses. At an interim analysis, subjects are considered censored if they are known to be uninfected and at risk at that point in time. Information on these subjects will continue to accrue through the surveillance period.

Algorithm 4-1: Pocock's test. k refers to the k^{th} interim analysis, K refers to the maximum number of interim analyses planned, z_k refers to the test statistic at the k^{th} interim analysis, and $c(K, \alpha)$ refers to the nominal significance level which is a function of K and α, the maximum type I error allowed.

```
for k = 1, ..., K do
  if |z_k| ≥ c(K, α) then
    stop, reject H_0
  else
    if k == K then
      stop, accept H_0
    else
      continue
    end if
  end if
end for
```

4.5 Epidemiological Model

To estimate the attack rate encountered by subjects in a given clinical trial—a key component for our cost/benefit calculations—we require information about the spread of the COVID-19 epidemic in the U.S. We use the Susceptible-Infected-Resolving-Dead-ReCovered with social distancing (SIRDC-SD) model proposed by Fernandez-Villaverde and Jones [117] because it is able to fit both the cumulative and daily number of deaths in all the states well despite being a simple model, to establish a baseline for the epidemic.

We assume that there is a constant population of N people and denote the number of people who are susceptible to infection, infected, resolving their infected status, dead, and recovered as S_t, I_t, R_t, D_t, and C_t, respectively.

$$N = S_t + I_t + R_t + D_t + C_t \tag{4.12}$$

Under SIRDC-SD, the dynamics of the epidemic are governed by the following differential equations:

$$\frac{dS_t}{dt} = -\frac{\beta(t) S_t I_t}{N} \tag{4.13}$$

$$\frac{dI_t}{dt} = \frac{\beta(t) S_t I_t}{N} - \gamma I_t \tag{4.14}$$

$$\frac{dR_t}{dt} = \gamma I_t - \theta R_t \tag{4.15}$$

$$\frac{dD_t}{dt} = \delta \theta R_t \tag{4.16}$$

$$\frac{dC_t}{dt} = (1-\delta)\theta R_t \tag{4.17}$$

Unlike most epidemiological models that assume a static contact rate β, the SIRDC-SD model assumes a contact rate parameter, $\beta(t)$, that decreases exponentially over time at a rate of λ from an initial value of β_0 to β^*:

$$\beta(t) = \beta_0 e^{-\lambda t} + \beta^*(1 - e^{\lambda t}) \tag{4.18}$$

This dynamic $\beta(t)$ incorporates the belief that social distancing over time will lead to a lower contact rate. This is particularly true in the U.S. where many cities have issued stay-at-home orders. Many people are also voluntarily wearing masks and are avoiding crowded places, which serve to reduce the contact rate.

The model assumes that infections resolve at a Poisson rate γ, which implies that a person is infectious for a period of $1/\gamma$ on average. Thereafter, the individual will stop being infectious and transition into the "resolving" state. Resolving cases will clear up at a Poisson rate of θ. There is an implicit assumption that people who recovered from the virus gain immunity to the virus and cannot be reinfected.

We estimate the model for each of the 50 states in the U.S. and Washington, D.C. using the time series of deaths in the U.S. obtained from the John Hopkins Center for Systems Science and Engineering (CSSE) COVID-19 repository [118, 119] as of June 16, 2020. See Appendix C.2 for our parameter estimation method.

To predict the path of the epidemic after the lockdowns are relaxed and/or vaccines are developed, we propose the Susceptible-Infected-Resolving-Dead-ReCovered-Vaccinated with social distancing (SIRDCV) model as an extension of the SIRDC-SD model which accounts for vaccination. See Appendix C.3 for details.

We consider three different scenarios for the evolution of the epidemic over time. In the first, we assume that the current situation will continue indefinitely until the end of the epidemic ("status quo"). That is, stay-home orders and bans on social gatherings will be extended until there are no new infections. We simply forecast ahead of time using the estimated parameters in this scenario.

In the second, we consider a partial reopening with strict monitoring across all states starting from June 15, 2020 ("ramp"). To model this, we assume a ramp function for $\beta(t)$ that increases over 90 days to 0.22 and remains at that level until the end of the epidemic. The parameters are chosen to imply a final R_0 of 1.1, which reflects close monitoring and contact tracing, and if needed, temporary quarantines to arrest clusters of infections that may pop up.

In the third, we consider the behavioral-based response proposed by John Cochrane ("behavioral"), whereby people voluntarily reduce social contact when they perceive

danger (e.g., when they observe that there is an uptick in the daily number of deaths) and increase social contact when they observe that there is a decrease in risk (e.g., when they observe a reduction in the daily number of deaths) [120]. See Appendix C.4 for the functional forms of β in the three scenarios.

Lastly, we assume that vaccines will be immediately available for distribution and inoculation upon licensure. This reflects how leading vaccine companies have been scaling up their manufacturing capabilities and producing millions of doses at industrial scale in parallel to the clinical trials [121, 122] and well before the demonstration of vaccine efficacy and safety, i.e., at-risk manufacturing. We model three ways that the susceptible population will be vaccinated upon vaccine licensure: 1M, 10M, and infinite doses administered per day. In the last case, the entire U.S. population is assumed to be vaccinated the day after licensure. While unrealistic, this gives an upper bound on the potential benefit of a vaccine approval.

4.6 Cost/Benefit Analysis Framework

We apply cost benefit analysis to quantify and compare the net value of each trial design. We focus on public health outcomes—that is, the risks of mortality and morbidity—and provide a qualitative discussion of the societal and financial impact in Section 4.8.

As shown by Montazerhodjat et al. [123], Isakov et al. [124], and Chaudhuri et al. [125], the value associated with a pathway can be decomposed into an in-trial cost and a post-trial benefit. The former measures the cost of conducting the study to volunteers in the trial while the latter estimates the net benefit of the trial to society at large:

$$\text{Net Value} = \text{Post-trial Benefit} - \text{In-trial Cost} \tag{4.19}$$

We quantify the cost of a trial design in terms of the number of COVID-19 infections and deaths observed in the clinical study. For post-trial benefit, we first

consider a baseline scenario in which a vaccine is never developed and the epidemic is allowed to run its course. Next, we simulate the case where a vaccine is approved at some point in time depending on the duration of the trial design. The post-trial benefit is then the difference in the cumulative number of infections and deaths in the population between the two scenarios, i.e., the incremental number of infections and deaths prevented with a vaccine licensure.

In our simulations, we consider a vaccine candidate with some efficacy ϵ and assume that infections in the clinical study follow a stochastic process (e.g., binomial distribution). Due to this randomness, false rejections of the efficacious vaccine might occur. This is also known as type II error. The false negative rate depends on the trial design (e.g., sample size, surveillance period, maximum type I error, and superiority testing) and the epidemiological model (e.g., attack rate in the clinical study). In cases where the vaccine candidate is rejected, net value will be negative since post-trial benefit is zero but cost has been incurred for conducting the clinical trial. Lastly, we assume that the hypothetical vaccine candidate is generally well tolerated and any vaccine-related adverse reactions are mild and negligible with respect to in-trial costs and post-trial benefits [126, 127, 128].

4.7 Results

We compute the expected net value of different trial designs using Monte Carlo simulations and asymptotic distributions of the efficacy test statistics (see Section 4.4). Fig. 4-1 illustrates the inputs, computations, and outputs of our simulation framework. We assume that all trials start on August 1, 2020, and simulate the epidemiological models until December 31, 2022. We perform sensitivity analysis over a wide range of trial design, epidemiological model, and population vaccination schedule assumptions (see Table 4.3), covering 756 different scenarios. We summarize our results in Table 4.4 and Appendix C.6. In addition to our results, we release an open-source version of our simulation software, and encourage readers to rerun our simulations with their own preferred set of assumptions and inputs.

Assuming superiority testing and a vaccine efficacy of 50%, we estimate the date of licensure of the hypothetical vaccine candidate to be some time in November 2021 under an RCT (476 days), between June and August 2021 under an ORCT (326 to 380 days), between April and June 2021 under an ARCT (246 to 306 days), and between March and June 2021 under an HCT (221 to 311 days). For specificity, we report estimated times to licensure using calendar dates and provide the corresponding number of days in parentheses. However, our simulations do depend on calendar dates in one respect: the epidemiological model used to estimate the attack rates depends on current data. Therefore, the estimates reported here are all based on extrapolated conditions as of August 1, 2020, and may need to be revised for other start dates.

Apart from an RCT which has a fixed trial duration, the dates of licensure from the ORCT and ARCT depend largely on the status of the epidemic during the clinical trial. If the transmission rate of the disease is low (e.g., due to social distancing or other non-pharmaceutical interventions), an extended surveillance period is required to accrue enough natural infections in order to observe a statistically significant difference in infection risk between the treatment arm and the control arm. Conversely, when the transmission rate is high, a short surveillance period is sufficient to observe significant results. We note that an HCT, on the other hand, does not depend on the epidemic situation but is instead limited by the time required to set up the challenge model. In general, we find that the time to licensure under ORCT and ARCT decreases with increasing vaccine efficacy: the greater the efficacy, the easier it is to observe a significant treatment effect.

We find that the ARCT provides the greatest expected net benefit among the three RCT designs in almost all scenarios. The utility of an HCT versus the RCTs, however, depends critically on the set-up time and the dynamics of the epidemic. For example, assuming superiority testing, a vaccine efficacy of 50%, the behavioral epidemiological model, and a population vaccination schedule of 10M doses per day, we estimate that the ARCT can help accelerate licensure by almost 8 months versus the RCT, thus preventing approximately 2.9M incremental infections and 23,000 incremental deaths

Table 4.3: Sensitivity analysis with respect to trial design, epidemiological model, and population vaccination schedule assumptions. The total number of configurations simulated is 756 (computed as the product of the last column).

Parameter	Values	Number of Combinations
Trial design	RCT, ORCT, ARCT, HCT (30-day set-up), HCT (60-day set-up), HCT (90-day set-up), HCT (120-day set-up)	7
Vaccine efficacy of hypothetical candidate (%)	30, 50, 70, 90	4
Efficacy requirement	Superiority, superiority by margin of 30% [89], superiority by margin of 50%	3
Epidemiological scenario	Status quo, ramp, behavioral	3
Population vaccination schedule (doses/day)	1M, 10M, infinite	3

from COVID-19 in the U.S. versus the latter.

Under the same set of assumptions, an HCT that requires 30 days to set up can further reduce the time to licensure by a month, thus preventing approximately 1.1M more infections and 8,000 more deaths versus the ARCT. However, the advantage of the HCT vanishes when its set-up time is long: an HCT that requires 90 days to set up takes about one month longer to reach licensure as compared to the ARCT, leading to around 1.0M more infections and 8,000 more deaths versus the latter (see Fig. 4-3a). Under such circumstances, the use of an HCT is worthwhile only when the prevalent transmission rate is low. If we consider the status quo scenario instead of the behavioral epidemiological model, the time to licensure is about one month shorter under the HCT than under the ARCT even with a 90 day set-up period (see Fig. 4-3b). In this case, the HCT prevents approximately 60,000 incremental infections and 500 incremental deaths versus the ARCT. We observe similar trends under superiority-by-margin testing at a threshold of 50%.

4.8 Discussion

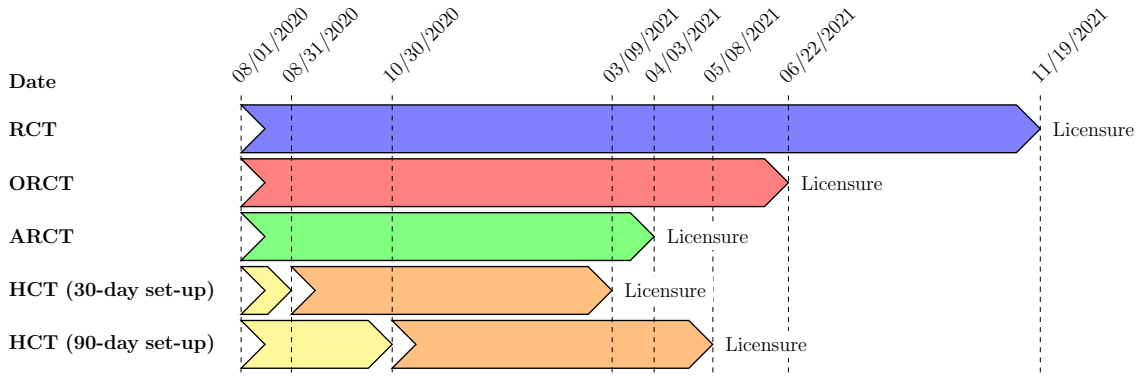
There has been a plethora of papers highlighting various ethical considerations for conducting HCTs [129, 130], some specifically for COVID-19 [98, 131, 132, 133, 134, 135]. Some of the main ethical concerns are: (1) what is the explicit scientific rationale

Table 4.4: Expected number of incremental infections and deaths avoided in the U.S. under different trial designs, vaccine efficacies, and epidemiological scenarios, assuming trials start on August 1, 2020, superiority testing, and 10M doses of a vaccine per day are available after licensure, compared to the baseline case in which no vaccine is ever approved.

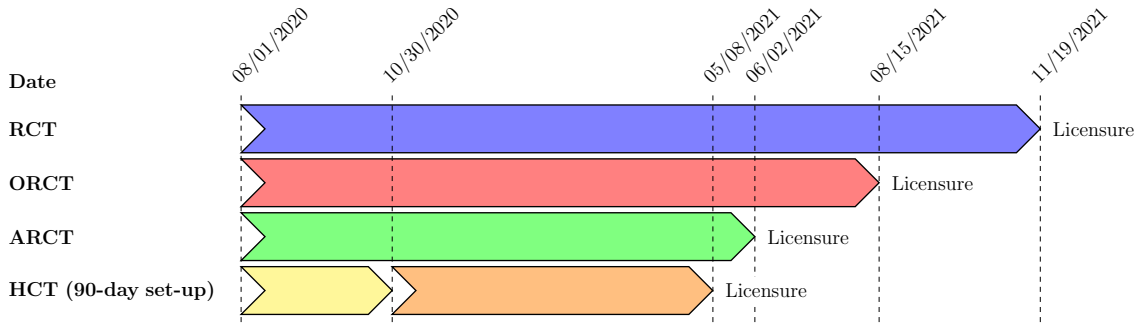
	Vaccine Efficacy (%)			
	30		**50**	
	$\mathbb{E}[\Delta\text{Infections}]$	$\mathbb{E}[\Delta\text{Deaths}]$	$\mathbb{E}[\Delta\text{Infections}]$	$\mathbb{E}[\Delta\text{Deaths}]$
Status Quo				
RCT	3,914	31	11,539	92
ORCT	5,589	45	16,802	134
ARCT	9,596	76	31,473	250
HCT (30-day set-up)	140,731	1,124	152,263	1,216
HCT (60-day set-up)	110,046	879	118,937	950
HCT (90-day set-up)	86,466	690	93,370	745
HCT (120-day set-up)	68,213	544	73,611	587
Behavioral				
RCT	363,382	2,845	386,081	3,026
ORCT	1,139,585	9,061	1,377,157	10,955
ARCT	2,588,881	20,647	3,248,449	25,924
HCT (30-day set-up)	3,903,566	31,167	4,309,316	34,411
HCT (60-day set-up)	2,795,316	22,301	3,082,676	24,598
HCT (90-day set-up)	2,011,244	16,028	2,211,985	17,633
HCT (120-day set-up)	1,466,239	11,668	1,605,833	12,784
Ramp				
RCT	1,075,634	8,316	1,131,531	8,764
ORCT	2,853,202	22,569	3,839,945	30,432
ARCT	5,711,310	45,401	7,442,922	59,253
HCT (30-day set-up)	8,744,377	69,672	9,452,413	75,330
HCT (60-day set-up)	6,814,762	54,235	7,381,425	58,762
HCT (90-day set-up)	5,266,925	41,851	5,711,663	45,404
HCT (120-day set-up)	4,053,134	32,141	4,396,033	34,879

Table 4.4 (continued): Expected number of incremental infections and deaths avoided in the U.S. under different trial designs, vaccine efficacies, and epidemiological scenarios, assuming trials start on August 1, 2020, superiority testing, and 10M doses of a vaccine per day are available after licensure, compared to the baseline case in which no vaccine is ever approved.

	Vaccine Efficacy (%)			
	70		**90**	
	$\mathbb{E}[\Delta\text{Infections}]$	$\mathbb{E}[\Delta\text{Deaths}]$	$\mathbb{E}[\Delta\text{Infections}]$	$\mathbb{E}[\Delta\text{Deaths}]$
Status Quo				
RCT	19,130	151	21,557	170
ORCT	33,757	269	50,288	401
ARCT	66,641	531	83,522	665
HCT (30-day set-up)	156,885	1,254	159,876	1,277
HCT (60-day set-up)	122,482	979	124,777	997
HCT (90-day set-up)	96,111	768	97,886	782
HCT (120-day set-up)	75,747	605	77,132	615
Behavioral				
RCT	397,396	3,117	404,562	3,174
ORCT	1,426,014	11,345	1,457,500	11,598
ARCT	3,389,541	27,052	3,473,035	27,720
HCT (30-day set-up)	4,481,448	35,789	4,591,750	36,671
HCT (60-day set-up)	3,205,159	25,579	3,283,975	26,209
HCT (90-day set-up)	2,297,350	18,316	2,352,436	18,757
HCT (120-day set-up)	1,664,613	13,255	1,702,601	13,558
Ramp				
RCT	1,160,564	8,996	1,179,234	9,145
ORCT	3,973,769	31,501	4,050,013	32,111
ARCT	7,924,650	63,107	8,071,866	64,285
HCT (30-day set-up)	9,725,022	77,511	9,897,591	78,892
HCT (60-day set-up)	7,602,878	60,534	7,743,514	61,659
HCT (90-day set-up)	5,887,421	46,811	5,999,381	47,706
HCT (120-day set-up)	4,532,400	35,970	4,619,521	36,667



(a) Under the behavioral epidemiological model.



(b) Under the status quo epidemiological model.

Figure 4-3: Dates of licensure under RCT, ORCT, ARCT, HCT (30-day set-up time), and HCT (90-day set-up time), assuming superiority testing, a vaccine efficacy of 50%, and a population vaccination schedule of 10M doses per day.

for, and societal value of, an HCT; (2) whether the risks of harm to the subjects and the public at large are understood by the scientists and have been minimized; (3) whether informed consents have been obtained from subjects after they are given full disclosures of the risks involved; and (4) whether the subjects have been selected fairly and given appropriate compensation for both the risk and actual harm brought on by HCTs. Most bioethicists generally accept that these concerns can be addressed within the existing ethical framework for human medical research.

Our work addresses the first of these ethical concerns by quantifying the relative cost/benefits of a COVID-19 HCT compared to traditional vaccine development pathways. We also discuss scientific justifications for HCTs by considering how conducting them can allow companies to estimate protection curves and correlates of protection, which can potentially accelerate future COVID-19 vaccine development by enabling immunogenicity trials.

However, our analysis does not address the remaining ethical considerations as they concern the execution of HCTs, which is beyond the scope of this study. Nonetheless, companies and scientists seeking to perform HCTs, and especially regulators, will have to address those concerns to preserve public trust and avoid a public backlash that could jeopardize other important medical research critical to addressing the current epidemic.

Some scientists argue that "a single death or severe illness in an otherwise healthy volunteer would be unconscionable." [135] However, it can be argued that allowing tens of thousands of individuals to die by denying the consent of an informed individual to take a calculated risk is equally unconscionable. In this study, we adopt the Benthamite approach [136], where every individual's utility is weighted equally in the aggregate utility function, as is the common convention in public economics analyses. Within this ethical perspective, our calculations show that an HCT can potentially provide substantial public health benefits in terms of accelerating vaccine development and reducing the burden of coronavirus-related mortality and morbidity in the U.S.—in some cases, by more than 1.1M infections and 8,000 deaths compared to the best performing non-challenge RCT—when conducted early in the pandemic's

life cycle and in cases where the spread of COVID-19 in the population is muted due to non-pharmaceutical interventions.

We also expect the financial costs of an HCT—even after accounting for the cost of liability protection—to be lower than those of a traditional vaccine efficacy RCT, adding further support for a challenge design (see Appendix C.5 for further discussion). While we have focused on public health outcomes here, it is clear that accelerated vaccine development provides tremendous societal and economic benefits as well—e.g., savings in insured medical costs, direct medical expenditures, and hospitalization costs, and accelerated economic recovery from an earlier reopening.

We emphasize that the expected costs and benefits of a clinical trial depend critically on many assumptions about existing conditions. For example, recruiting subjects in sufficient numbers and diversity can sometimes present a challenge for clinical trials involving experimental vaccines. (Although, in the case of HCTs for COVID-19, the organization 1Day Sooner reports over 32,000 registered volunteers as of July 27, 2020 [137].) Also, we do not include set-up time for non-challenge RCTs because phase 3 vaccine efficacy trials are already imminent as of now. Moreover, we assume a relatively short set-up time for HCTs because a challenge study can be set up relatively quickly using a wild-type strain [108], and the National Institute of Allergy and Infectious Diseases (NIAID) appears to have already made some headway in manufacturing challenge doses [138, 139, 140]. If, instead, we assume comparable set-up times (e.g., two months) and start dates for both an HCT and non-challenge RCTs, we expect that an HCT can accelerate licensure by two months when compared to an adaptive RCT (assuming superiority testing, a vaccine efficacy of 50%, and the behavioral epidemiological model). Some have argued that at least one to two years is required to develop a robust model from scratch [135]. In this case, our results indicate that an ARCT will almost always be faster than an HCT. However, even if an HCT with a long set-up time does not lead to faster vaccine licensures over an ARCT given current conditions, the creation of a standing HCT agent and setting up an HCT now can provide a hedge against potential failures in the current crop of vaccine candidates. By having an approved, ready-to-go challenge virus and ready-to-go

HCT sites that vaccine developers can access immediately, the approval process for as-yet-untested SARS-CoV-2 vaccine candidates can be accelerated when required. For a pandemic like COVID-19, such a hedge will almost always show substantial net benefits relative to its costs.

HCTs have several other benefits that will be more obvious as the pandemic progresses. They require many fewer eligible volunteers, whose numbers will dwindle as the pandemic progresses. They do not depend on attack rates at clinical trial sites which are notoriously difficult to estimate and highly dependent on non-pharmaceutical interventions such as lockdowns and other social-distancing policies. They also avoid logistical problems such as identifying subjects, obtaining subjects' consent, obtaining institutional review board's approval or tracking subjects, particularly when attempting large-scale clinical trials in places where contract research organizations (CROs) have little experience.

It is conceivable that multiple vaccines—instead of the single vaccine in our simulation study—are tested concurrently in a single trial design [141]. For example, five vaccines, such as those selected by Operation Warp Speed [142], could be tested concurrently in a six-arm trial (five vaccine arms and a control arm), requiring 40% fewer test subjects, thereby reducing in-trial expected morbidity and mortality costs by the same amount. The benefits can be increased if an adaptive platform clinical trial—designed to eliminate ineffective vaccines at the first signs of futility—is adopted. A clinical trial testing multiple vaccines can also reduce competition for volunteers, a problem that continues to plague vaccine developers [143].

We choose to quantify the cost and benefits of the clinical trials by measuring the number of infections and deaths avoided, and refrain from performing a traditional health technology assessment, such as comparing the economic value of an HCT versus an RCT using quality-adjusted life years measures or willingness to pay estimates such as the value of a statistical life. Performing such computations is straightforward given the output of our simulations, but we have refrained from doing so in deference to non-economist stakeholders who find it offensive to use any pecuniary measures when discussing the loss of human life.

Finally, our analysis focuses mainly on the U.S. for practical reasons involving access to data with which to calibrate our simulations and the broader goal of informing U.S. public health officials and policymakers as the country enters the final stages of vaccine development. However, we note that vaccine companies such as Pfizer/BioNTech are also looking at recruitment in the Southern Hemisphere (e.g., Brazil [144]), which can affect the rate at which these events accumulate in trials, depending on the spread of COVID-19 in those countries. In addition, a vaccine licensure may apply internationally. Given that the U.S. currently comprises 25% of all confirmed COVID-19 cases (as of July 7, 2020) [118], if the assumptions made in our study also hold internationally, the net benefits for all the clinical trials will scale by a factor of 4, in which case HCTs can save an additional 4.4M infections and 32,000 deaths compared to the best performing RCT in certain situations.

We highlight that these figures depend heavily on the development of the epidemic in the U.S. moving forward. We have considered three simple scenarios, status quo, ramp, and behavioral, corresponding to low transmission, moderate transmission, and behavioral-based response, respectively. There are clearly many other sources of uncertainty that are not reflected here. For example, non-adherence to social distancing advisories and/or resistance to precaution recommendations such as wearing a mask in public will lead to an uncontrolled outbreak, which will help to accelerate non-challenge RCTs, making them attractive even when compared to an HCT with a short set-up time. We have found it difficult and impractical to incorporate these uncertainties in our assumptions due to the speed at which things are evolving and the unpredictability of public reaction. In addition, studies that have attempted to incorporate such uncertainties in their epidemic model report huge error bounds in their projections [145]. The wide confidence intervals prevent us from drawing any useful conclusions, which severely limit the usefulness of such models. Therefore, we recommend readers not to take our results as final or definitive, but to re-run our simulations with their own preferred set of assumptions, calibrated using the most current epidemiological data.

4.9 Conclusion

In this chapter, we present a systematic, transparent, reproducible, and principled way to quantify the trade-offs between the different COVID-19 vaccine efficacy clinical trial designs under various scenarios. We hope that this framework will allow stakeholders such as vaccine developers, policymakers, and HCT volunteers to understand the implications of their actions (or inaction). Our results also contribute to the moral and ethical debate about HCTs amidst this crisis. One of the main ethical and regulatory barriers to the acceptability of HCTs has been the lack of immediately available, effective "rescue therapies." However, this situation is already changing with the emergency use authorizations of remdesivir and convalescent plasma therapy. Our findings, coupled with these therapeutics, may have an impact on how critical public health decisions like whether to employ HCTs are made.

Part III

New Business Models

Chapter 5

Impact of University Technology Licensing: A Case Study of MIT

Academic institutions play an increasingly critical role in the biotechnology industry through technology licensing and the creation of startups, but there is little data on their performance and the magnitude of their impact. In this chapter, we address the knowledge and data gap through a systematic study of technology licensing by the Massachusetts Institute of Technology (MIT). Using data on the 76 therapeutics-focused life sciences companies formed through MIT's Technology Licensing Office from 1983 to 2017, we construct several measures of impact including MIT patents cited in the Orange Book, capital raised, outcomes from mergers and acquisitions, patents granted to MIT intellectual property licensees, drug candidates discovered, and U.S. drug approvals, a key benchmark of innovation in the biopharmaceutical industry. As of the cutoff of our dataset, we find four small molecule drugs that cited MIT patents in the Orange Book. However, we find that MIT licensees played a directly traceable role in the approval of 31 drugs by the U.S. Food and Drug Administration (excluding candidates acquired after phase 3) from 1991 to 2017, of which 55% were a new molecular entity or new biological entity, and 55% were granted priority review, an indication of addressing an unmet medical need. Our methodology provides a useful framework for other academic institutions to track the outcomes of their intellectual property in the therapeutics domain.

5.1 Introduction

The process of drug development in the pharmaceutical industry is undergoing a profound shift in its industrial organization. Increasingly, smaller biotechnology firms apply recent academic research in the life sciences to develop new drugs, which are then acquired by pharmaceutical giants using their financial war chests and access to low-cost capital to purchase expertise [146].

Technology licensing is a key driver of this process in the United States. Prior to the passage of the Bayh-Dole Act of 1980, few universities in the U.S. held significant portfolios of patent rights. The Bayh-Dole Act was explicitly intended to promote the commercialization of products developed from federally funded research. Under the Act, grantees of federal funds, such as universities, would be allowed to pursue patent rights to their research—rights that, under the terms of many earlier federal grants, had been automatically assigned to the government.

The passage of Bayh-Dole was contemporaneous with an enormous expansion of patents granted to universities. In 1979, the year before the act's passage, only 264 U.S. patents were awarded to universities in the United States; by 1997, this number had grown to 2,436 [147]. In 2016, American universities were issued 7,021 U.S. patents and had filed for 2,507 patents abroad [148].

More importantly for the biopharma sector, the number of startup companies formed as a result of technology licensing increased from 145 in 1994 to 278 in 2000 to 1,024 in 2016 [148, 149]. And this trend has had a remarkable impact on drug development, as Tables 5.1 and 5.2 show. Of the 30 top-selling drugs worldwide in 2000, only five were traceable to universities, and only two of these were in the top 10; the remaining 25 were developed by big pharma. By 2015, more than one-third of the top 30 drugs were sourced from academia, and 60% of the top 10.

Despite the growing importance of technology licensing to the biomedical ecosystem, there has been surprisingly little data collected on the impact of technology transfer by academia. We hope to change this deficit by providing a detailed analysis of the portfolio of life sciences intellectual property (IP) of the Massachusetts Institute

of Technology (MIT) from 1983–2017. Although other studies have been published on patenting activity at specific universities [150], to the best of our knowledge, this study is the first systematic analysis of a portfolio of therapeutics companies that have licensed the IP of a specific academic institution.

Under the prior leadership of Lita Nelsen, and now Lesley Millar-Nicholson, the MIT Technology Licensing Office (TLO) has for decades played a critical role in transferring IP from MIT to private enterprise in Kendall Square and beyond. MIT's biomedical research engine has contributed significantly to the growth of the biotech ecosystem that has emerged in the Boston/Cambridge area. The life sciences sector—which includes firms developing therapeutics, medical devices, diagnostics, and research tools—is an active area of innovation at MIT.

In this chapter, we limit our scope to a subset of the life sciences sector, one that includes biotechnology, pharmaceuticals, and other therapeutics, defined as firms in the business of developing new drugs, either small molecules or biologics. We focus on therapeutics for two reasons. First, the ecosystem physically surrounding MIT has its primary focus on drug discovery and development. Second, this focus allows us to use the number of drug approvals as a core metric of innovation, as is done every year by journals such as Nature Reviews Drug Discovery [151].

We develop a framework for tracking innovation originating in academia that builds upon prior work measuring innovation in the pharmaceutical industry [146, 152, 153]. Using the Orange Book, we attempt to link MIT IP to drugs approved by the U.S. Food and Drug Administration (FDA) [154, 155]. However, explicit lines of causality are rare due to the complexity of the journey from IP licensing to FDA approvals, which complicates the determination of the degree of contribution from the IP versus the degree of contribution from the company.

To provide a clearer picture of the impact of MIT IP, we separately track the origin of each drug in our dataset to weight the contributions of MIT licensees—that is, companies that have licensed MIT IP—to their respective approved drugs. To measure different types of innovation, these drugs are also labeled as new molecular entities (NMEs) or new biological entities (NBEs), and as drugs granted priority review (PR)

Table 5.1: Top 30 worldwide top-selling drugs in 2000. The last column indicates the university/hospital that was involved in the drug's discovery or early stage development [155, 156, 157, 158, 159]. Blanks indicate that no university/hospital was involved.

Marketer	**Drug**	**Worldwide Sales**	**University/Hospital**
AstraZeneca	Prilosec	6,260	
Merck	Zocor	5,280	
Pfizer	Lipitor	5,030	
Pfizer	Norvasc	3,361	
TAP Pharmaceuticals	Prevacid	2,740	
Johnson & Johnson	Procrit	2,709	University of Chicago
Pfizer; Pharmacia	Celebrex	2,614	Brigham Young University
Eli Lilly	Prozac	2,585	
Eli Lilly	Zyprexa	2,366	
GlaxoSmithKline	Paxil	2,349	
Schering Plough	Claritin	2,194	
Merck	Vioxx	2,160	
Pfizer	Zoloft	2,140	
Amgen	Epogen	1,963	University of Chicago
Wyeth	Premarin	1,870	
GlaxoSmithKline	Augmentin	1,847	
Merck	Vasotec	1,790	
Bristol-Myers Squibb	Pravachol	1,766	
Bristol-Myers Squibb	Glucophage	1,718	
Merck	Cozaar	1,715	
Johnson & Johnson	Tylenol	1,680	
Novo Nordisk	Novolin	1,671	
Bayer	Cipro; Ciprobay	1,648	
Johnson & Johnson	Risperdal	1,603	
Bristol-Myers Squibb	Taxol	1,561	Florida State University
Pfizer	Zithromax	1,382	
Schering Plough	Intron A	1,360	University of Zurich
Pfizer	Viagra	1,344	
Pfizer	Neurontin	1,334	
GlaxoSmithKline	Flixotide; Flovent	1,334	
	Total ($M)	69,374	10,207
	Percent of Total		15

[146, 152, 153]. Ultimately, this framework enables us to compare the innovation contributed by MIT licensees against the benchmark of historical pharmaceutical industry averages. In addition, we analyze the following company outcomes: capital raised in initial public offerings (IPOs), mergers and acquisitions (M&A) volumes, total drug candidates developed, and patent granted. Finally, we examine several case studies that demonstrate the broader contribution of MIT to the growth of the biotech industry beyond IP licensing, e.g., the role of MIT faculty as co-founders and advisors to biotech companies that have gone on to be very successful.

Table 5.2: Top 30 worldwide top-selling drugs in 2015. The last column indicates the university/hospital that was involved in the drug's discovery or early stage development [155, 156, 160, 161, 162, 163]. Blanks indicate that no university/hospital was involved.

Marketer	**Drug**	**Worldwide Sales**	**University/Hospital**
Abbvie; Eisai	Humira	14,359	Rockefeller University; Scripps
Gilead	Harvoni	13,864	University of Heidelberg; Rockefeller University
Amgen; Pfizer; Takeda	Enbrel	9,037	Massachusetts General Hospital; University of Texas
Johnson & Johnson; Merck; Mitsubishi Tanabe Pharma	Remicade	8,151	New York University
Pharmstandard; Roche	Rituxan	7,393	
Sanofi	Lantus	7,089	
Roche	Avastin	6,945	
Roche	Herceptin	6,794	University of California, Los Angeles
Daewoong Pharmaceutical; Pfizer	Prevnar 13	6,328	
Celgene	Revlimid	5,801	Boston Children's Hospital
GlaxoSmithKline	Seretide; Advair	5,625	
AstraZeneca	Crestor	5,381	
Gilead	Sovaldi	5,276	
Pfizer	Lyrica	4,876	Northwestern University
Amgen	Neulasta	4,800	Memorial Sloan-Kettering Cancer Center
Novartis	Gleevec	4,658	Dana-Farber Cancer Institute; Oregon Health & Science University
Bayer; Regeneron	Eylea	4,372	
Teva	Copaxone	4,029	Weizmann Institute of Science
Boehringer Ingelheim; Eli Lilly	Spiriva	3,942	
Bayer; Johnson & Johnson	Xarelto	3,930	
Merck	Januvia	3,870	Tufts University
Novartis; Roche	Lucentis	3,639	
Biogen	Tecfidera	3,638	
Gilead	Truvada	3,567	Emory University; Yale University
AstraZeneca	Symbicort	3,394	
AstraZeneca; Pfizer	Nexium	3,202	
Bristol-Myers Squibb; Gilead; Merck	Atripla	3,134	Emory University; Yale University
Novo Nordisk	NovoRapid; NovoLog	3,082	
Bristol-Myers Squibb; Otsuka	Abilify	2,896	
Eli Lilly	Humalog	2,842	
	Total ($M)	165,914	86,940
	Percent of Total		52

Table 5.3: Summary of MIT portfolio of life science and therapeutics companies.

	Count
Private Companies	43
Public Companies	
IPO	
Alive	10
Acquired	13
Bankrupt	3
Reverse Merger	
Alive	4
Others	
Alive	2
Bankrupt	1
Total	76

5.2 Data

From an initial list of 225 life sciences MIT licensees, we identify 76 therapeutics companies. (See Appendix D.1 for more details.) We further narrow the bulk of our analysis to 33 companies that were or are currently publicly listed on an exchange, due to the availability of detailed information in their financial filings. Table 5.3 provides an overview of the MIT portfolio of companies reviewed as part of our analysis.

With the exception of Aprecia Pharmaceuticals, none of the private licensees have brought a drug to market singlehandedly. This is not surprising given the capital-intensive process of commercializing therapeutic candidates. To address these capital needs, biotech companies will tap the public markets, or engage in strategic business development or M&A transactions with larger biopharmaceutical companies. Thus, we will focus primarily on public companies, but also summarize private M&A outcomes.

5.3 Measures of Impact

5.3.1 Orange Book Citations

Following the approach in Stevens et al. [155], we use the Orange Book [154, 164, 165] and the United States Patent and Trademark Office (USPTO) databases [166, 167] to link MIT IP and FDA-approved drugs (cutoff at December 2017). To identify drugs

that owe their origin, at least in part, to MIT IP, we search the publication for New Drug Applications (NDAs) that cite patents assigned to MIT (see Appendix D.2).

We find four small molecule drugs that cited MIT patents (see Table 5.4). (See Appendix D.2 for citations that fall outside the scope of our analysis or occurred after our cutoff.) Redux was brought to approval by Indevus in partnership with Wyeth for obesity. However, its approval was later withdrawn due to safety concerns [168]. Gliadel, Sarafem, and Spritam are examples of a direct link between MIT IP and approval. The technology behind Gliadel, an implantable wafer loaded with the chemotherapy agent carmustine for use in glioblastoma treatment, was invented by Robert Langer's lab at MIT, and developed by Guilford Pharmaceuticals, a spinout of Scios Nova. Scios acquired Nova Pharmaceutical, which had licensed the IP from MIT. Scios Nova was uninterested in developing the product, so MIT facilitated the creation of Guilford, a spinout started specifically to develop and commercialize the wafer and related technologies [169].

In the development of Sarafem, MIT professor Richard Wurtman patented his discovery that low serotonin levels in the brain contributed to premenstrual dysphoric disorder (PMDD). He then founded Indevus to license the patent to Eli Lilly, which already marketed Prozac, a selective serotonin reuptake inhibitor that increased serotonin levels in the brain for the treatment of depression. Eli Lilly subsequently developed the compound for PMDD and launched a newly branded version called Sarafem [170].

The use of Orange Book citations as a measure of impact of MIT IP has several limitations. First, the absence of a medical school and hospital predisposes research at MIT to platform-based technology that may be applicable to many different drugs/diseases, over optimization of specific drug compounds. Therefore, most of MIT patents do not cover composition of matter, which are most pertinent for market exclusivity protection and inclusion in the Orange Book; they usually focus on explication of mechanisms. This also means that MIT licensees are likely to spend a large part of the patent term seeking the optimal target for the platform, e.g., Sangamo and its zinc finger nuclease (ZFN) gene-editing platform. Given that the time to

Table 5.4: List of MIT IP citations for small molecule drugs in the Orange Book. See Appendix D.2 for citations that fall outside the scope of our analysis or occurred after our cutoff.

Company	Drug	Patent No.	Title
Indevus	Sarafem	4,035,511	Process for promoting analgesia
Indevus	Sarafem	4,083,982	Process for producing analgesia
Indevus	Sarafem	4,971,998	Methods for treating the premenstrual or late luteal phase syndrome
Indevus	Redux	4,309,445	d-Fenfluramine for modifying feeding behavior
Guilford	Gliadel	4,757,128	High molecular weight polyanhydride and preparation thereof
Guilford	Gliadel	4,789,724	Preparation of anhydride copolymers
Guilford	Gliadel	5,179,189	Fatty acid terminated polyanhydrides
Aprecia	Spritam	6,471,992	Dosage form exhibiting rapid disperse properties, methods of use and process for the manufacture of same
Aprecia	Spritam	9,463,160	Dosage form exhibiting rapid disperse properties, methods of use and process for the manufacture of same

market for new drugs can be as long as 20 years, it is not uncommon for MIT patents to expire before a drug is approved.

To further complicate the link between academic IP and approvals, companies often license additional IP from other institutions, and build upon existing technology to file new patent applications during the development process. In addition, investigational drugs typically undergo many iterations of formulation studies. Consequently, it is not surprising if the initial MIT IP is ultimately displaced from the Orange Book by more recent patents that can afford greater protection.

Clearly, the impact of MIT IP extends beyond citations in the Orange Book. MIT IP also plays an important role in catalyzing a company's financing by attracting interest from venture capitalists, and serving as foundation for future research and development (R&D). To provide a clearer picture of the innovation contributed by MIT IP, we analyze the financials and the R&D portfolios of MIT licensees, and provide several other measures of impact including capital raised, M&A outcomes, drug candidates discovered, drug approvals, and patents granted.

5.3.2 Initial Public Offerings

Of the 33 public biotech MIT licensees, 26 completed the standard IPO process, four reverse-merged into publicly listed companies, and three listed through alternative pathways. We use the Form S-1 financial filings—registration forms submitted to the

Securities and Exchange Commission (SEC) for new securities—to produce Table 5.5, which summarizes the results of the IPOs that have sufficiently available data. (The financial filings for firms that conducted an IPO prior to 1996 were not available.) It includes the proposed share price range, the final share price, the proceeds net of underwriting fees, the shares issued and outstanding, and the dilution due to financing.

On average, the IPOs of MIT licensees raised about $41M in net proceeds, diluted the existing shareholders 26% in the offering, and achieved a post-IPO valuation of $218M. Given that therapeutics companies typically conduct an IPO to finance R&D and clinical trials when their assets are under development, the average valuation post-IPO is in line with expectations. The capital generated by the IPOs and the resulting shareholder dilution were also plotted over time (see Appendix D.3). Over the studied period, MIT licensees raised more capital, though this also came with increased dilution to shareholders. After adjusting for inflation using the Consumer Price Index (CPI) and the Biomedical Research and Development Price Index (BRDPI), we find that net proceeds experienced a high level of growth, an indication of the dramatic expansion of the biotech capital markets over the past two decades.

Table 5.5: Available IPO data from financial filings of 26 publicly traded therapeutics companies. The column Priced compares the final share price to its corresponding proposed range. It provides an assessment of the demand for the company's equity, and thus in some sense measures the "success" of the IPO. *Abbreviations*: adj, adjusted.

							Net Proceeds (US$M)[9]					IPO Valuation (US$M)		
Company	**Date**	**Low (US$)**	**High (US$)**	**Final (US$)**	**Priced[1]**	**Shares Issued (Mil)**	**Nominal**	**CPI Adj (2017$)**	**BRDPI Adj (2017$)**	**Shares Outstanding (Mil)**	**Dilution (%)**	**Nominal**	**CPI Adj (2017$)**	**BRDPI Adj (2017$)**
Scios Nova[2]	Jan-83			12.0		1.0	*12*	*30*	*41*					
Cistron	Jun-86			1.0		5.0	*5*	*11*	*14*					
Interneuron[3]	Mar-90			2.0		5.3	*11*	*20*	*25*	14.9	36	30	56	71
Alkermes	Jul-91			10.0		1.8	16	29	37	7.3	24	73	132	166
Cell Genesys	Jan-93	9.0	11.0	11.0	1	4.0	41	69	85					
Arris[4]	Nov-93			7.0		2.5	16	28	34	8.4	30	59	100	123
ARIAD	May-94			8.0		1.9	14	23	27	15.7	12	125	207	252
Guilford[5]	Jun-94			8.0		1.9	13	21	26					
Millennium	May-96	11.0	12.0	12.0	1	4.5	50	78	95	22.8	20	274	428	519
T Cell[6]	Jun-96			12.0		2.3	25	39	47	19.9	12	239	374	453
Algos	Sep-96	14.0	16.0	14.0	-1	3.5	46	71	86	15.5	23	218	340	412
Cubist	Oct-96	6.0	7.0	6.0	-1	2.5	14	22	26	9.1	27	55	86	104
Abgenix[7]	Jul-98	10.0	12.0	8.0	-2	2.5	19	28	33	10.6	24	85	128	152
Sangamo	Apr-00	15.0	17.0	15.0	-1	3.5	49	70	81	20.9	17	313	445	521
Praecis	May-00	10.0	12.0	10.0	-1	8.0	74	106	124	40.0	20	400	569	665
Acusphere	Oct-03	13.0	15.0	14.0	0	3.8	49	65	74	14.3	26	200	266	301
Alnylam	Jun-04	6.0	8.0	6.0	-1	5.0	28	36	41	19.3	26	116	150	168
Momenta	Jun-04	6.5	7.0	6.5	-1	5.4	32	42	47	24.6	22	160	207	232
Tengion	Apr-10	8.0	10.0	5.0	-2	6.0	28	32	33	12.4	49	62	69	72
Merrimack	Apr-12	8.0	10.0	7.0	-2	14.3	96	102	106	92.4	15	647	691	720
bluebird	Jun-13	14.0	16.0	17.0	2	5.9	94	99	103	22.8	26	388	408	424
Conatus	Jul-13	10.0	12.0	11.0	0	6.0	61	65	67	15.6	39	171	180	187
BIND	Sep-13	14.0	16.0	15.0	0	4.7	66	69	72	15.8	30	237	249	259
Cerulean[8]	Apr-14	11.0	13.0	7.0	-2	8.5	56	58	60	19.0	45	133	138	142
Editas	Feb-16	16.0	18.0	16.0	-1	5.9	88	90	90	35.7	17	571	583	586
Selecta	Jun-16	14.0	16.0	14.0	-1	5.0	65	66	67	17.9	28	251	256	257
Mean		10.9	12.7	9.8	-1	4.6	41	53	59	21.6	26	218	276	308
Median		10.5	12.0	10.0	-1	4.6	37	50	53	16.8	25	186	228	254

[1] -2: Priced IPO below initial range; -1 = Priced IPO at the bottom of initial range; 0 = Priced IPO at the middle of initial range. [2] Now Scios. [3] Now Indevus. [4] Now Axys.
[5] Scios Nova spin-off. [6] Now Celldex. [7] Cell Genesys spin-off. [8] Now Dare. [9] *Italics*: estimated from gross proceeds.

5.3.3 Mergers and Acquisitions

Among the 76 biotech MIT licensees, 23 companies were acquired in 11 private and 12 public M&A transactions, with a total value of $30.7B (see Table 5.6 and Appendix D.4). This large volume was driven primarily by three public company deals totaling $23.5B. Millennium (acquired by Takeda), ARIAD (acquired by Takeda), and Cubist (acquired by Merck) were all fully integrated biopharmaceutical companies with potential blockbuster assets at the time of transaction. Millennium marketed Velcade for multiple myeloma and mantle cell lymphoma, but also had a diverse pipeline of small molecule inhibitors and monoclonal antibodies. ARIAD marketed Iclusig for select hematological malignancies, and had potential blockbuster brigatinib for ALK+ non-small cell lung cancer on the cusp of FDA approval. Cubist marketed a variety of antibiotics, including blockbuster Cubicin for S. aureus and complicated skin and skin structure infections, and had a late-stage clinical pipeline of antibiotics.

A significant difference in size between private and public M&A is expected, due to the crucial role that public markets play in funding the growth of development-stage biotech companies [171]. However, one particularly notable private deal was Merck's acquisition of SmartCells, founded by Todd Zion, an MIT chemical engineer. In 2010, SmartCells was sold for over $500M, including upfront and downstream milestones, based on a preclinical asset called MK-2640 for type-1 diabetes [172, 173]. The value created for SmartCells shareholders is unknown, as the breakdown of upfront cash and contingent value rights was not disclosed. Another large private deal was Civitas, which was on the cusp of an IPO when it sold to Acorda Therapeutics for $525M. Civitas was developing a drug for Parkinson's disease ready for phase 3, and had filed an S-1 just one month before Acorda stepped in [174].

5.3.4 Research and Development Pipeline

The average drug takes at least a decade of translational research and several clinical studies before it is approved by the FDA. Each phase of clinical testing costs millions of dollars, typically leading to multiple publications and a better understanding of

Table 5.6: Acquisition values of MIT biotech companies. *Abbreviations*: adj, adjusted.

Type	MIT Company	Acquirer	Date	Value (US$M)		
				Nominal	CPI Adj (2017$)	BRDPI Adj (2017$)
Private	Enzytech	Alkermes	Dec-92	30	52	65
Private	Oculon	Pharmos	Apr-95	4	6	8
Private	Peptimmune	Genzyme	Jul-99			
Private	Sirenade	NeuroTek AG	Dec-05	9	11	13
Private	FoldRx	Pfizer	Sep-10			
Private	SmartCells[1]	Merck	Dec-10	>500	>562	>580
Private	Surface Logix	Nano Terra	Apr-11			
Private	Pervasis[1]	Shire	Apr-12	200	214	223
Private	Ceregene[2]	Sangamo	Aug-13	>1	>1	>1
Private	Synglyco	Coden	Nov-13			
Private	Civitas	Acorda	Sep-14	525	544	562
Total				1,269	1,391	1,450
Public	Algos	Endo	Nov-99	241	355	417
Public	Cistron[1]	Celltech	Apr-00	25	36	42
Public	Axys	Celera	Jun-01	173	239	279
Public	Scios	Johnson & Johnson	Apr-03	2,400	3,197	3,616
Public	Guilford[3]	MGI	Jul-05	178	223	248
Public	Abgenix[2]	Amgen	Dec-05	2,200	2,761	3,076
Public	Praecis	GlaxoSmithKline	Dec-06	55	67	73
Public	Millennium	Takeda	Apr-08	8,800	10,019	10,823
Public	Indevus[1]	Endo	Jan-09	637	728	761
Public	Cell Genesys	Biosante	Jul-09	38	43	45
Public	Cubist	Merck	Dec-14	9,500	9,836	10,162
Public	ARIAD	Takeda	Feb-17	5,200	5,200	5,200
Public	Cerulean	Dare	Jul-17	20	20	20
Total				29,467	32,725	34,762

[1] Acquisition value includes contingent value rights. [2] Cell Genesys spin-off. [3] Scios spin-off.

the targeted disease and its pathway. Because of this complexity, we believe that examining the full R&D pipelines of MIT licensees—that is, all drugs developed over a company's lifespan—can provide valuable insights about their economic and biomedical contribution to the biopharma industry.

In order to track their R&D pipelines, we went through each company's annual financial filings to the SEC (Form 10-K), and manually extracted all investigational compounds in the clinical phase. We focus on the number of unique pipeline drug candidates (quantity), the highest stage of development for any indication (depth), and the therapeutic areas involved (breadth), with less emphasis given to the specific number of indications targeted. This is to maintain consistency with Mullard [151], which excludes label expansions, but it is also due to the difficulties in quantifying the scope of a candidate during its early clinical phases. (See Appendix D.5 for details.)

The final dataset consists of 281 drug candidates from 33 public MIT licensees over 23 years of development, spanning from 1994–2017 (Figs. 5-1 and 5-2; see Appendix D.5 for breakdown by company). We find that oncology, neurological, and genitourinary-related drugs are the most popular areas of development in MIT biotech companies—over 50% of the companies were involved in at least one of these therapeutic areas. Of the 281 candidates, 51 are FDA-approved products. Among the companies, Alkermes outperforms the rest in both quantity and depth, leading the group with 22 approved products and an aggregate of 46 pipeline drugs developed or marketed in its portfolio. Other notable companies include Millennium with 26 compounds, and Indevus with 21 drug candidates.

We note that the numbers here include drugs that were acquired by MIT licensees after FDA approval, and fail to capture the impact of MIT IP that led to FDA-approved products following their acquisition by other companies. To provide a more accurate picture of the biopharma innovation in the MIT portfolio, in the next section, we categorize the approved drugs in the dataset to better reflect their origins and the contributions of MIT licensees.

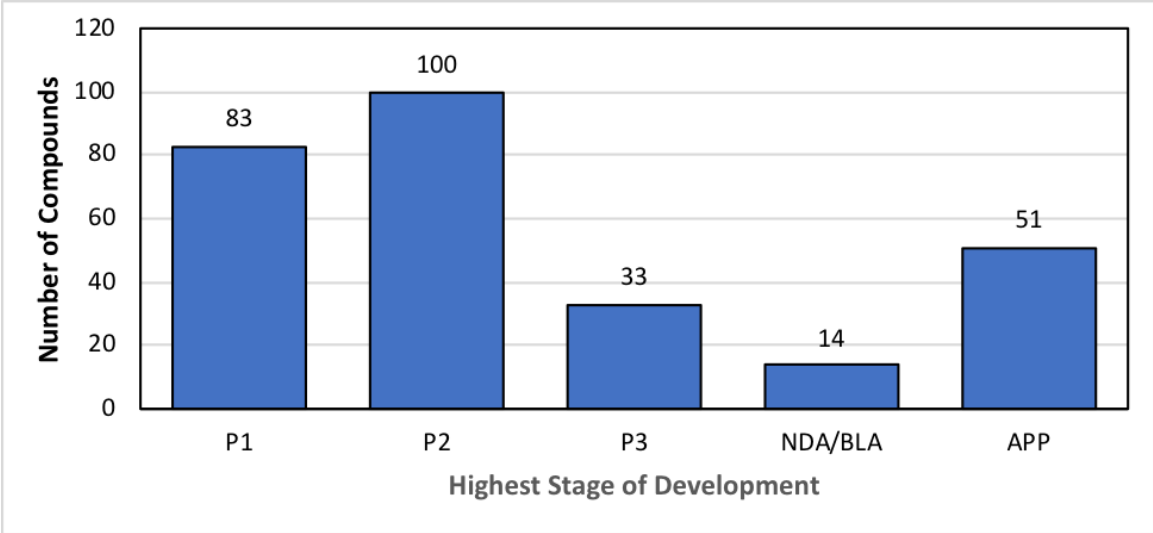


Figure 5-1: Highest stage of development of pipeline candidates of MIT licensees. *Abbreviations*: BLA: biologics license application; APP: approval.

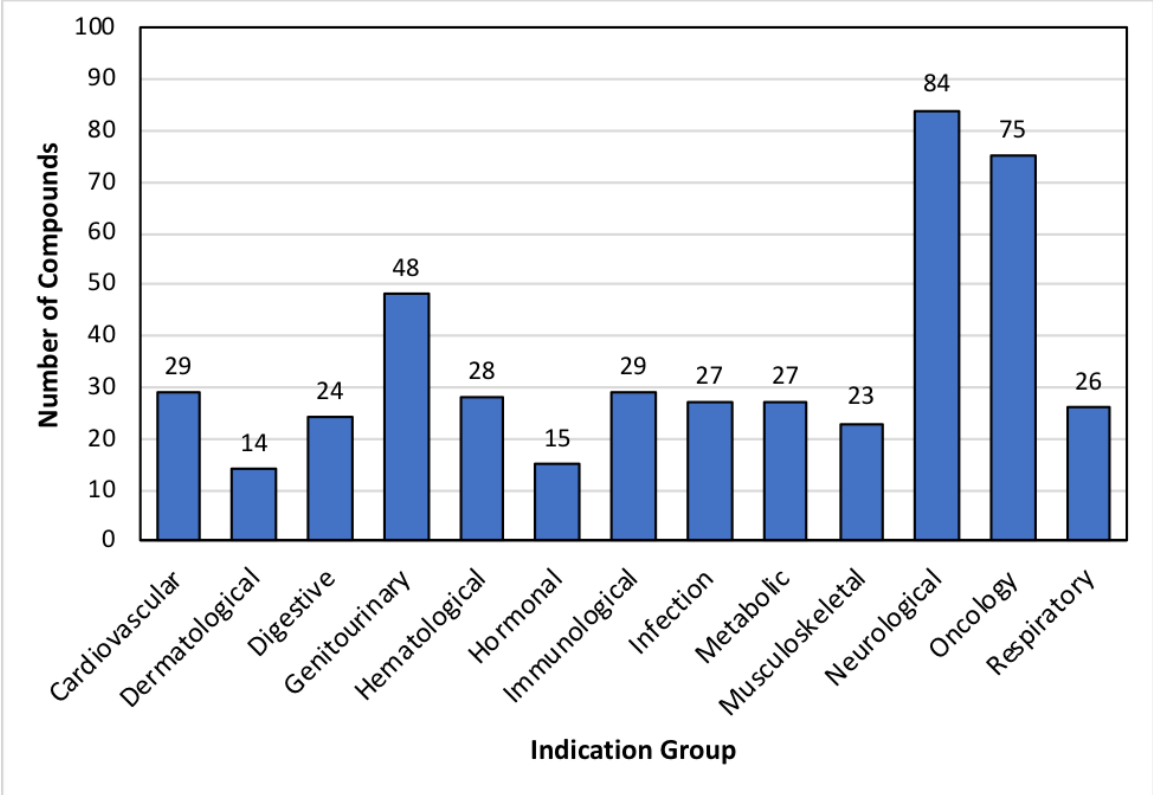


Figure 5-2: Indication groups of pipeline candidates of MIT licensees.

5.3.5 Drug Approvals

The number of drug approvals by the FDA is the key benchmark of success in the biopharma industry. However, the number of approvals of NMEs/NBEs and of drugs with a PR designation can help further differentiate novel drugs and drugs that address significant unmet medical needs or provide substantial clinical benefits [146, 151, 175, 176]. To capture both types of innovation, we screen the financial filings of the 33 public companies, and use the Drugs@FDA database to generate a list of approved drugs with their targeted indications, years of approval, properties, NME/NBE statuses, and PR designations (see Appendix D.6).

Although label expansions for already approved drugs were excluded earlier to maintain consistency with the prior literature [151], they are accounted for separately in Appendix D.6. For inclusion in our dataset, we define a label expansion as the FDA approval of a drug for a different disease than prior approvals, and omit expansions into various lines of therapy or patient populations within a disease. For example, in 2003, Velcade received initial approval to treat third-line multiple myeloma. We exclude the 2005 expansion of Velcade into second-line multiple myeloma, but we do include the 2006 addition of mantle cell lymphoma to Velcade's label.

The journey from MIT-licensee company formation and IP licensing to a successful drug approval is rarely straightforward. The high cost, high risk, and lengthy timeline of drug development often results in tortuous paths, and can involve split ownership in the form of licensing deals and collaboration agreements, or the exchange of ownership over time, often through M&A transactions [146]. Therefore, we classify drug approvals in four ways to more accurately identify the origins of the assets and clarify the contributions of the MIT licensees to their respective FDA-approved drugs: Partners, Originators, Originators*, and Acquirers (see Fig. 5-3).

Partners led the clinical development of drugs that were initially discovered by MIT licensees or were enabled by technologies from MIT licensees. Partners were common in the dataset, as biotech firms often license their technology to validate their science, generate non-dilutive capital, or share risk. For example, Curis discov-

ered a hedgehog pathway inhibitor for basal cell carcinoma that eventually became Erivedge under a collaboration agreement granting Genentech an exclusive, royalty-bearing license. Though Curis played a key role in the discovery of the compound, Genentech had complete responsibility over all clinical development and commercialization, ultimately spending significantly more resources and time on the asset. In contrast, ARIAD discovered and advanced Iclusig through clinical trials on its own without any partners.

The Originator classifier indicates ownership of assets at varying stages of development that were ultimately acquired and developed by MIT licensees. For example, Millennium acquired LeukoSite in 1999, which came with Campath and LDP-341. LDP-341 later became Velcade, a first-in-class proteasome inhibitor for multiple myeloma. In contrast to the case of ARIAD, where the approved molecules originated within company labs, Millennium did not contribute to the discovery of Velcade. The Originator classifier highlights the cases in which any MIT IP was likely far removed from the discovery of the assets.

Originators* differentiates drugs that were acquired after phase 3, which includes compounds that were NDA-ready, NDA-filed, or already marketed. When Millennium acquired LeukoSite, an NDA for Campath in chronic lymphocytic leukemia (CLL) had already been filed, while LDP-341 was only in phase 1. Thus, Millennium contributed meaningfully to the clinical development of Velcade, but not Campath. The MIT licensee objectively had no role in the development of Originator* drugs, let alone any contribution from MIT IP. These drugs are excluded from our analyses.

Many successful MIT licensees were bought by "Acquirers." For example, ARIAD, Abgenix, and Millennium had pipeline assets that were eventually approved by the FDA following their acquisition. In these cases, we still attribute credit to the MIT licensees for their involvement. For example, Takeda acquired Millennium in 2008 and retained the business as its dedicated U.S. oncology subsidiary. In 2015, Millennium and Takeda brought Ninlaro, a second-generation proteasome inhibitor, to market [177]. We consider Ninlaro as developed by Millennium with Takeda as a "Partner" to attribute credit appropriately, instead of excluding the asset entirely because the

Table 5.7: Summary statistics of drug approvals by MIT licensees.

		NME/NBE		PR		NME/NBE/PR	
	Total	**n**	**%**	**n**	**%**	**n**	**%**
All	58	28	48	25	43	34	59
All (excluding Originators*)	31	17	55	17	55	22	71
Drug development led by an MIT company (excluding Originators* and Partners)	16	9	56	9	56	12	75

approval occurred after Millennium was no longer a standalone entity. Similarly, Abgenix's XenoMouse monoclonal antibody technology continued to produce drug candidates like Repatha [178] after the company was acquired by Amgen. We list such cases in Appendix D.6 and include them in our subsequent analysis.

Including seven drugs approved post-acquisition, a total of 58 unique drugs have been on the market and under the ownership of MIT licensees (see Table 5.7). However, these 58 include 27 Originator* drugs that were acquired by MIT licensees, which dilute the impact of MIT IP on compounds for which MIT licensees played a role in their discovery or clinical development. Excluding these assets, 22 out of 31 (71%) were an NME/NBE or had PR, key indicators of innovation and impact on unmet needs. Further excluding assets classified as "Partners" narrows the group to those drugs for which MIT licensees led their development to market. This subset had 16 approvals, of which 12 (75%) were an NME/NBE or had PR. These final 16 include a number of well-known drugs, including Gliadel, Vivitrol, Aristada, Cubicin, Onivyde, and Velcade.

5.3.6 Intellectual Property

In the context of evaluating the impact of MIT IP, the number of patents granted to MIT licensees could be used as a proxy for the contribution of MIT's licensees to biopharma innovation. However, it is important to note that this metric is not ideal. The mission of the biopharma industry is ultimately to improve the health outcomes of patients. It is more appropriate to measure innovation by assessing the approval of drugs rather than granting of patents, whose technology may never reach patients.

The USPTO database was used to collect data for the 33 public MIT licensees

Year of Approval	Number of Drugs							
	All	Special Status		Classifications				
		NME/NBE	PR	None	Partner	Originator	Originator*	Acquirer
<1990	1						1	1
1990	1				1		1	
1991	2				2		1	
1992								
1993								
1994								
1995								
1996	5	3	1	1	3		3	2
1997								
1998	2	1	2				2	1
1999	2	1	2		2		1	
2000								
2001	3	2	1	1	1		2	1
2002	2				2		2	
2003	5	4	4	1	2	2	1	2
2004	3	1			1	1	2	2
2005	4	1	1	1	2	1	2	1
2006	2	1	2	1	1			1
2007	2			1			1	2
2008	3	2			2		3	1
2009	1				1		1	
2010	2	2	2		2		1	1
2011	1	1	1				1	1
2012	3	2	2	1	2			1
2013	1				1	1		
2014	4	2	2	1	1	2	1	3
2015	7	3	3	2	3	3	1	3
2016								
2017	2	2	2		2			1
Total	**58**	**28**	**25**	**10**	**31**	**10**	**27**	**24**

Figure 5-3: Summary statistics of special status and classifications of FDA-approved drugs with MIT licensee ownership or contribution. The 58 drugs include seven which were approved post-acquisition of MIT licensees. The "partner" classification applies if another firm led clinical development of the drug enabled by the MIT company's technology. The "originator" classification applies if the drug was acquired by the MIT company. The "originator*" classification applies if the drug acquired by the MIT licensee was a post-phase 3 asset (NDA-ready, NDA-filed, or marketed). The "acquirer" classification applies if the MIT company was the lead developer but was later acquired. The "none" tag applies if the MIT licensee was both the originator and lead developer of the approved drug. For example, there were two FDA approvals in 1991 that are associated with MIT licensees. The clinical development of both approvals was led by partners, i.e., both tagged with "partner" classification. One of the assets were acquired post-phase 3 by an MIT licensee for development, i.e., one tagged with "originator*" classification. See Appendix D.6 for the list of approvals.

(cutoff at December 2017; see Appendix D.7). The 33 companies licensed 258 unique patents from MIT in initial startup agreements totaling $39.9M in royalties. These companies were additionally granted 2,512 patents between 1985–2017, clearly showing that MIT licensees continue to innovate beyond an initial license from MIT.

The number of patents filed by a therapeutics company can be very different depending on the nature of the innovation involved. While certain types of innovation lend themselves to large patent estates to protect an investigational compound, others might rely on just a few patents based on core biological insights. We find that companies that developed novel technology platforms, e.g., Millennium, Alnylam, and Sangamo, were the most productive in terms of number of patents grated per year. In contrast, companies focused on developing in-licensed assets, such as Conatus Pharmaceuticals, tend to be focused on clinical development rather than on innovating new technology. In general, we do not find a strong correlation between the number of drugs approvals and the number of patents granted for MIT licensees (see Appendix D.7).

Although patent data does not measure the direct impact on patients, it does reveal one aspect of MIT's contributions not reflected in drug approvals. Companies such as Alnylam and Sangamo are pioneering new technologies, and had not yet achieved FDA approval before the cutoff of our dataset. However, as suggested by their extensive patent portfolios both firms contributed significant advances to the scientific community's understanding of siRNA and ZFN technologies as therapeutics during this time. (In August 2018, Alnylam received FDA approval for patisiran, a first-in-class siRNA therapeutic [179].)

5.4 Discussion

An analysis of the 2,529 new drugs approved by the FDA from 1991–2017 shows that 31% were NMEs, and 24% had PR (see Appendix D.8). In contrast, MIT licensees played a traceable role in the approval of 31 drugs over the same time period, of which 55% were NMEs and 55% had PR. This comparison is limited because of its small

sample size, and because the preference for NMEs and PR candidates by smaller biotech companies is unknown. Nevertheless, this suggests that MIT licensees may have been more innovative than the industry average.

The convoluted link between academic IP and FDA approval makes an analysis that tracks the outcomes of academic IP difficult. This limitation arises due to multiple factors. IP licensing is only an early step in a lengthy drug development timeline, companies may license IP from several institutions, and asset ownership is frequently multi-partied and variable over time. However, we believe that our approach—examining Orange Book citations, IPOs, M&As, R&D pipeline, drugs approvals, and intellectual property of MIT licensees—provides a fair and comprehensive framework to simultaneously acknowledge MIT origins and recognize the complex, multi-party contributions that extend beyond MIT and are required to commercialize drugs.

In some cases, MIT licensees found success after pivoting away from its initial IP and strategy. For example, Cubist Pharmaceuticals purchased daptomycin, a discontinued compound for Gram-positive infections in phase 2, from Eli Lilly in 1997 [180]. Cubist pushed daptomycin across the finish line as Cubicin in 2003. The drug subsequently became a blockbuster, and enabled Cubist to further acquire Adolor, Optimer, Calixa, and Trius, each of which resulted in an FDA-approved drug. Other examples include ARIAD and Millennium.

However, even without clear links between drug approvals and academic patents, academic IP still contributes to three critical aspects of a company's success: bringing together unique people and talent, catalyzing a company's financing, and serving as a foundation for future R&D. The case of Millennium is a prime example: Its technology platform enabled the firm to sign multiple partnership deals with large pharmaceutical companies and raise significant capital from the public markets. By 1998, Millennium had struck deals totaling $1B with pharmaceutical giants Eli Lilly, Roche, and Bayer [181]. These alliances provided Millennium with substantial funding and access to key downstream technologies for drug development. In 2000, Millennium raised over $1B from a follow-on public offering and a convertible debt sale. Its

lucrative partnerships validated its platform and attracted significant capital from investors. This combination of business development and financing activity ultimately allowed Millennium to make its transformative acquisitions of LeukoSite and COR Therapeutics. Clearly, the initial licensing of MIT IP had a broad impact on the organization.

Our analysis does not capture all aspects of MIT's contribution to therapeutic innovation. Within our dataset, biotech companies such as Alkermes, Abgenix, and Millennium have licensed their technologies to large pharmaceutical companies. Although drugs disclosed within the financial filings of MIT licensees were tracked, it is possible that others predominantly owned by outside partners have been missed. Specifics of early-stage drug development activity are not necessarily disclosed in a large pharmaceutical company's financial filings, and a biotechnology company may be contractually limited in its own disclosures of such programs. Similarly, access to public company financial filings prior to 1994 was limited, and preclinical assets or drug discovery technology of an acquired private company may remain undisclosed.

One example that falls outside of our framework is the case of Idun. Idun was founded in 1993 and licensed multiple MIT patents based on the work of Dr. Robert Horvitz. Idun and Abbott Laboratories (now AbbVie) collaborated to develop inhibitors of Bcl-2, a regulator of apoptosis. While the initial molecule produced by this collaboration did not progress due to poor oral bioavailability, Abbott developed follow-on compounds navitoclax and venetoclax. The latter has been approved for the treatment of various hematological malignancies including CLL. It has also been granted multiple breakthrough therapy designations based on its profound therapeutic impact particularly in diseases of high unmet need such as acute myeloid leukemia.

Further beyond our dataset, four biotech giants, Genzyme, Biogen, Amgen, and Genentech, were co-founded by MIT faculty or people with connections to MIT. However, these companies did not license MIT IP, and consequently were not included in our analysis [182, 183]. It is not uncommon for MIT faculty and affiliates to launch therapeutics companies without licensing MIT IP, such as Verastem, co-founded by Robert Weinberg and Eric Lander [184]. Finally, MIT research publications that are

not translated into patents also advance the therapeutics landscape. Our analysis does not capture these sources of impact because they do not involve an explicit licensing transaction with the MIT TLO.

On the other hand, as shown by the analysis of IP generated by the MIT licensees, entirely new classes of potentially transformative drugs are on the horizon whose development uses MIT IP. A brief survey of drugs in the development phase underscores MIT's contribution to the current creation of innovative therapeutics. Alnylam, bluebird bio, Editas Medicine, and Sangamo are just four examples of firms that have licensed MIT IP to develop platform technologies capable of discovering new drugs.

Alnylam was founded in 2002 by a group of MIT faculty, including Robert Langer and Phillip Sharp, to develop RNAi therapeutics based on siRNA discoveries. The firm licensed patents from MIT on the formulation and delivery of siRNAs, and also engaged in a five-year research collaboration to improve delivery to target tissues. Alnylam has a broad clinical development pipeline of seven assets, including patisiran, an RNAi therapeutic for hereditary ATTR amyloidosis, a severe neuropathy, under PR by the FDA as of the cutoff of our dataset [185]. (In August 2018, Alnylam received FDA approval for patisiran as a first-of-its kind targeted RNA-based therapy [179].) See Appendix D.9 for other examples.

It must be noted that other parties contributed to these innovative platforms in addition to MIT, such as Tekmira, which licensed its siRNA patents to Alnylam, and the Scripps Research Institute, which licensed its zinc finger proteins IP to Sangamo. In fact, Sangamo's most recent annual report references the ZFN IP licensed from the California Institute of Technology and the University of Utah, but not the 1996 patent license agreement between Sangamo and MIT. Although the original patent from 1996 may not be scientifically important relative to the more recent ZFN patents, perhaps it was critical to nucleate the company and enable Sangamo to become the leader of ZFN technology.

This highlights the critical point that developing a drug starting from an initial academic patent can require over 20 years, at which time the foundational patent

may be expired and no longer commercially relevant. Most of MIT licenses are not composition of matter patents, which are the most commercially valuable Orange-Book listable patents. Unlike many other institutions active in academic stage life sciences research, MIT does not have a medical school. This precludes large animal or early clinical testing prior to academic licensing and may bias the typical MIT license towards technology that underpin a new platform rather than towards an identifiable drug that would have a clearer, shorter path to market. A company that licenses technology may spend a decade developing the platform and identifying the optimal targets/diseases to pursue. So, it is not surprising if MIT patents have expired by the time a drug is approved. Nevertheless, MIT IP has played an important role in the formation of innovative companies and development of novel therapies. Numerous drug candidates produced by these companies are likely to reach the market in the coming years.

5.5 Conclusion

Due to the complexity of academic IP transactions and the difficulty in determining causation, the systematic tracking of the outcomes of academic IP transactions has been neglected, despite its obvious benefit to all players and its potential to increase the rate of biomedical innovation. The use of the framework developed in this chapter to analyze university IP would not only allow cross-university comparisons, but could provide convincing data in favor of increasing funding to bridge the financing gap for preclinical assets and drug discovery technology. As shown by the recent $130M collaboration between Deerfield Management and Duke University [186], there is both precedent and desire for such funding to accelerate innovation at the academic level.

More speculatively, with such data in hand, one could conceive of an "Academic Translational Medicine" (ATM) fund that raises money from limited partners to invest in therapeutics companies that license IP among a consortium of universities, each of which contributes IP to a commingled pool using a standardized IP agreement, managed by an external third-party portfolio manager dedicated to selecting and de-

veloping assets on behalf of investors in the fund, similar to the CRISPR patent pool initiative by MPEG LA [187]. If this fund were large enough, it could also provide additional value-added services such as animal studies, medicinal chemistry, toxicology, and even clinical-trial designs, not unlike the support offered by the National Institutes of Health's National Center for Advancing Translational Sciences.

Some seasoned biotech VCs may be skeptical about the prospect of multiple universities, each with its own unique group of stakeholders, goals, and constraints, collaborating productively in commercializing their IP. However, the downward pressure by funding organizations on overhead rates at a time of tremendous expansion in life sciences research has created the potential for a more conducive environment for cooperation between multiple universities, especially if motivated by the prospect of significant funding. One example is the research consortium between Celgene, the University of Pennsylvania, Columbia University, Johns Hopkins, and Mount Sinai, established to accelerate the discovery of new cancer treatments [188].

Given the assumption that innovation in biopharma is primarily constrained by a lack of capital rather than talent or worthwhile ideas, an ATM fund could greatly accelerate the commercialization of new drug discovery technologies and the discovery of new and innovative medicines. As suggested by our data, such an investment could prove beneficial not only for the researcher and the investor, but also patients.

Chapter 6

Financing Correlated Drug Development Projects

Current business models have struggled to support early-stage drug development due to the rising costs of clinical trials and a shift in research focus to more complex scientific pathways that have higher chances of failure. To address this issue, Fernandez et al. [23] proposed a novel megafund structure to finance early-stage translational research. Fagnan et al. [189, 190] applied the approach to a portfolio of therapeutics for orphan diseases. In this chapter, we extend the simulation framework proposed in previous studies to account for correlation between phase transitions in a portfolio of drug development projects for rare diseases, thus making the model a more realistic representation of biopharma research and development and also allowing us to evaluate the tail risks of the megafund more accurately. In addition, we update the parameters used by Fagnan et al. [190] with more recent probability of success (PoS) estimates. We find that the performance of the megafund becomes less attractive when correlation between projects is introduced. However, the risk of default and the expected returns of the vanilla megafund remain promising even under moderate levels of correlation. Furthermore, we find that a leveraged megafund outperforms an equity-only structure over a wide range of assumptions about correlation and PoS. Despite our focus on orphan drugs in this chapter, our framework can be easily generalized to arbitrary drug development portfolios.

6.1 Introduction

The drug development process has become increasingly expensive and risky over the past few decades. This phenomenon can be attributed to the rising cost of clinical trials and a shift in research focus to more complex biological mechanisms that are potentially more transformative but also have higher risks of failure. As a result, the current business model for research and development (R&D) in biopharma is becoming less effective. This is reflected in the decline of R&D productivity and the lackluster performance of investments in the biotech and pharma sectors in recent years.

Fernandez et al. [23] proposed a megafund structure to address this issue. This structure pools a large number of biomedical programs together in its portfolio, thus diversifying the risk of drug development and increasing the likelihood of success through multiple "shots on goal." By tranching this structure and redistributing the risk of default, the megafund can tap into the fixed income market, a substantially larger pool of capital than the conventional sources of biopharma R&D financing—public and private equity—but one traditionally unwilling to participate in biopharma investments due to the risky and fragmented nature of drug development. The megafund finances its large portfolio using capital raised from issuing equity and debt, i.e., bonds collateralized by the portfolio of pipeline drugs and their associated intellectual property. Simulation results by Fernandez et al. [23] show that this alternative financial structure can yield reasonable returns for investors in both types of securities.

More recently, Fagnan et al. [189, 190] have applied the megafund approach to early-stage drug development, the riskiest part of the drug discovery process, and the one where funding is also the scarcest. They found that the megafund structure is particularly well suited for financing orphan drugs, as orphan drugs typically have higher probabilities of success, lower clinical costs, and shorter development times than their non-orphan counterparts. In their simulations, an orphan drug megafund managed to generate double-digit annualized returns with a portfolio of only ten to twenty orphan drug projects.

In this chapter, we use the multi-state, multi-period simulation framework described in Fernandez et al. [23] and Fagnan et al. [189, 190] to analyze the potential performance of an orphan drug megafund. However, we note that the assumption of independent phase transitions of previous megafund studies rarely holds in practice, since drug candidates tend to exhibit some amount of correlation with one another depending on the similarities of their underlying treatment pathways. We demonstrate that the presence of correlated transitions has important consequences for the performance of the megafund, as seen in our formal derivation and empirical results (Section 6.2.4 and Section 6.3, respectively). To obtain a more realistic representation of biopharma R&D, we examine the use of a single-factor model with a Gaussian copula to model correlations between pipeline drugs in the portfolio. This approach allows us to evaluate the tail risks of the megafund more accurately.

In addition, we update the parameters for clinical trial durations and probabilities of success based on the estimates reported by Wong et al. [17] in a recent study. We also simulate the performance of several different megafund structures with correlated portfolios (vanilla, guarantee-backed, and equity-only), and perform a sensitivity analysis of several key parameters in our framework, specifically the capital structure, the portfolio acquisition strategy, the level of correlation between projects, and the probabilities of success for phase transitions.

Along with our results, we release an open-source version of our simulation software, and encourage readers to rerun our simulations with their own preferred set of assumptions and inputs.

6.2 Methods

6.2.1 Simulation Framework

A drug development megafund is a financial entity that pools and repackages a portfolio of pipeline drug assets into an arbitrary number of tranches with different risk, reward, and maturity characteristics. It offers the repackaged securities to investors

as "research-backed obligations" (RBOs)—that is, debt and equity securities backed by the pool of underlying drug assets—and uses the capital raised to finance the development of pipeline drugs in its portfolio. The RBO is structured to follow a strict priority for cash flow distributions. In general, senior debt tranches have first priority on the cash flows generated by the portfolio, and therefore have the best credit rating. Mezzanine tranches have the second claim on cash flows, but they are compensated by higher coupon rates for the higher risk of default. Finally, equity holders bear the risk of first loss, but at the same time, they are entitled to all the residual cash from debt repayment.

Here, we consider an RBO structure with the same three types of tranches: senior debt, junior debt, and equity. In addition to the subordination of cash flows, however, we adopt credit enhancement mechanisms designed to provide additional protection for the bondholders [25]: we maintain a reserve account at levels in excess of the fund's current liabilities, i.e., its short-term interest and principal payments. This account is tracked periodically to ensure that it remains above a minimum target level, failing which assets are liquidated to cover the shortfall. These coverage triggers can prevent the fund from abruptly going into default by identifying potential shortfalls ahead of time, thus giving portfolio managers sufficient lead time to monetize available assets. This is especially important for assets which are relatively illiquid, such as drug development programs, where extensive negotiation is necessary, and there is a lag between the sale of a project and the actual cash inflow.

We assume an investment structure based on the licensing framework commonly used in the biopharmaceutical industry [23]. The megafund first acquires a majority stake in each drug development program for an upfront payment. In our model, pipeline drugs undergo the standard drug approval process: starting from pre-clinical research, advancing through phase 1, phase 2, phase 3, and New Drug Application (NDA), before finally gaining approval. Each stage of development requires a certain amount of time and funding, at the end of which the program will either progress to the next higher phase or be discontinued. In exchange for the majority stake, the megafund is responsible for all clinical trial expenses ("development costs"), and also

any milestone payments due to the project investigators for the successful completion of each phase of development. Pipeline drugs are typically financed up to a specific target phase before being monetized, but they can also be sold for revenue at any point during development. As mentioned earlier, we assume that there is some lag time between a sale and the actual cash inflow.

The cash flow waterfall of the megafund is complex. In addition to periodic debt and interest payments, investments in pipeline drugs at each phase of development must be carefully managed to ensure that the interest coverage ratio remains above a minimum level. Depending on the performance of the portfolio, projects may need to be either put on hold until sufficient capital becomes available, or prematurely liquidated prior to the target phase, to make up for any shortfall in cash flows.

We use a multi-period Monte Carlo simulation model to evaluate the returns of the megafund over a fixed time horizon (see Fig. 6-1). We assume that the fund assembles a portfolio of drug development programs at the start of the simulation. We compute the financial statistics of the fund and the performance of the portfolio at discretized time steps ("periods"), and allocate the cash flows in each period according to the waterfall structure. We assume that the phase transitions follow a stochastic process, with the clinical trial cost, testing duration, and asset valuation drawn from continuous random distributions. When the end of life of the fund is reached, or if the fund defaults on its bond payments at any point during its tenor, the portfolio is liquidated. The proceeds are used to repay all outstanding debt, and any residual cash is distributed to the equity holders. Thereafter, the megafund is dissolved.

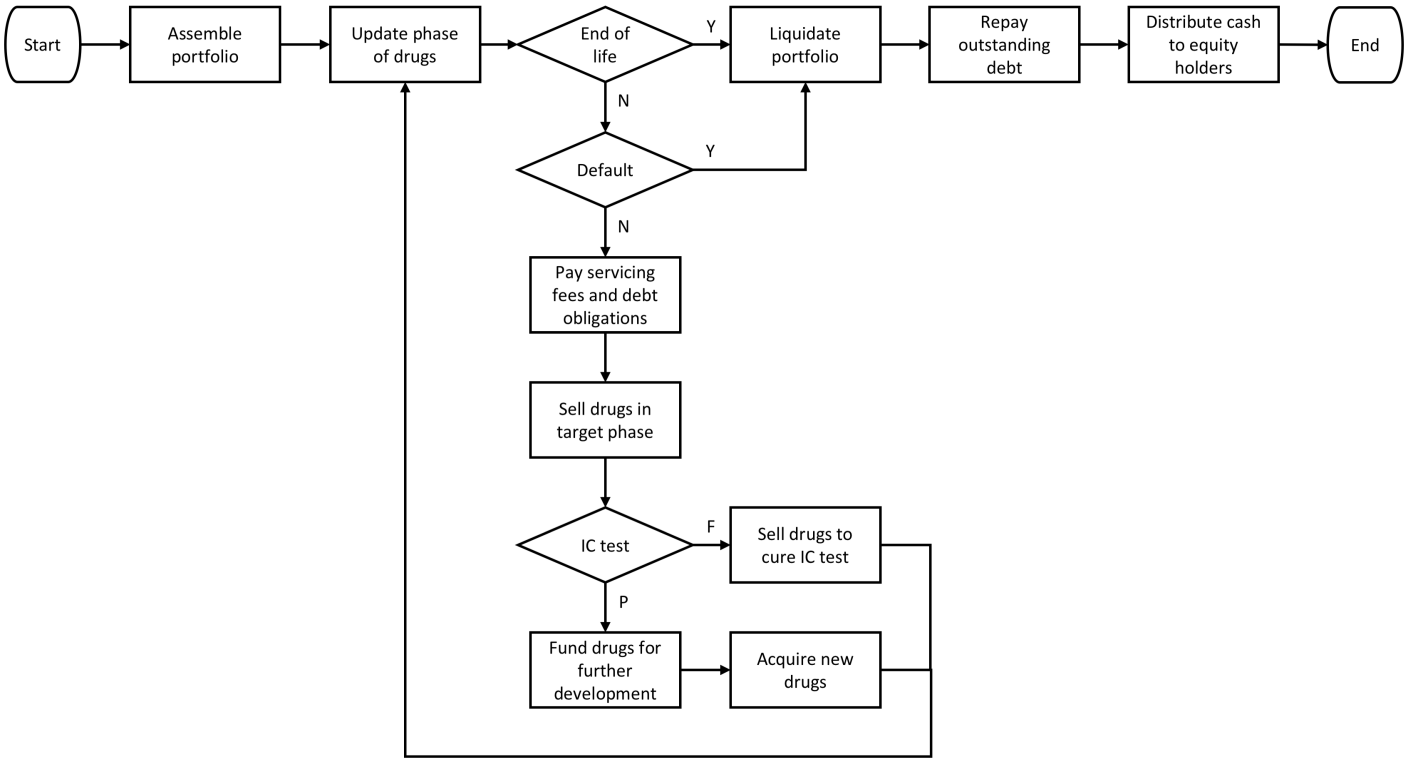


Figure 6-1: Simulation framework for the megafund. The cash flow waterfall incorporates the subordination of cash flow, the credit enhancement mechanism, the investment structure of drug development, and the drug approval process. Pipeline drugs that successfully advance to the next phase are funded only if there is sufficient cash remaining after settling current liabilities and fulfilling the interest coverage test. If not, they are held in the portfolio without further development until additional capital becomes available. The sale of pipeline drugs is the dominant source of cash flow for the megafund. *Abbreviations*: Y, yes; N, no; F, fail; P, pass; IC, interest coverage.

6.2.2 Parameters

In our simulation, we assume that the megafund has a tenor of ten years and a capital structure comprising three tranches: a senior debt tranche, a junior debt tranche, and an equity tranche. At the start of each simulation, the fund raises $575M of capital, including $250M from senior debt, $50M from junior debt, and $275M from equity. The senior bonds are structured to have a maturity of five years with an annual coupon rate of 5% and the junior bonds nine years with 8%. Each tranche is amortized evenly (i.e., straight-line amortization) over the four-year interval preceding its date of maturity. The schedule is structured so that principal payments do not overlap, and junior bonds are retired only after senior bonds have been fully redeemed.

We discretize the simulation horizon into six-month time periods, and assume that the megafund makes debt and coupon payments at the end of each period (i.e., semi-annual payouts) according to the amortization schedules of the bonds. In addition, we assume that the sale of each drug development program takes a year to settle, from the initiation of transaction to the receipt of cash proceeds. Therefore, we consider a simulation horizon of ten years, which spans the tenor of the fund, and leaves an additional year at the end for portfolio liquidation. In the absence of default, all clinical assets that have not already been sold or discontinued at the end of the ninth year are liquidated, and the proceeds received in the last period are distributed to the equity investors.

Here, we focus on early-stage orphan drug development projects for the RBO portfolio. We assume that the megafund acquires 23 pre-clinical programs at the start of simulation, with the aim of funding them through the completion of phase 2 clinical testing (i.e., to phase 3) before their sale. This is the maximum number of drugs the megafund can afford to finance, based on the amount of capital raised and the expected development cost required for each drug to reach the target phase. Each acquisition grants the megafund an 85% ownership stake in the asset, thus entitling the fund to the same portion of proceeds when the asset is monetized.

The simulation framework relies on several important modeling assumptions re-

garding the cost of clinical trials, the duration of clinical testing, asset valuation, and phase transition. Following the approach by Fagnan et al. [190], we model the cost and duration of clinical trials at each phase of development as independent and identically distributed (IID) log-normal random variables. We impose an upper bound on the development cost of each phase to limit the maximum possible expense that can be incurred per compound. Upfront costs and milestone payments are taken to be constants based on the phase of development. Similar to our treatment of development costs, we assume an upper-bounded log-normal distribution for drug asset valuation at each stage of development. However, instead of imposing independence, we introduce pairwise correlation of market valuation between projects using a single-factor model (see Section 6.2.3).

We model the drug development process as a sequence of Bernoulli trials, i.e., as a Bernoulli process: at each phase of development k, a pipeline drug has some probability p_k of advancing to the next higher phase $k+1$ ("success") and probability $1+p_k$ of being discontinued ("failure"). The time spent in each phase depends on the clinical testing duration drawn from the log-normal distribution described earlier. In our model, discontinuation is assumed to be an absorbing state, i.e., a drug that has been withdrawn can no longer reenter the development process (see Fig. 6-2).

Fagnan et al. [190] modeled phase transitions as IID random variables. However, we note that the assumption of independence rarely holds in practice, since drugs tend to exhibit some amount of correlation with one another, depending on the similarities in their underlying scientific pathways, mechanisms, and targets (e.g., two drugs with similar mechanisms of action are likely to have similar outcomes in testing). The presence of correlation has significant implications for the performance of the megafund. In general, correlation between assets introduces systematic risk to the portfolio, which by definition cannot be diversified away. Increased correlation leads to fatter tails in the distribution of the number of successful projects in the portfolio (see Section 6.2.4), which in turn adversely affects the credit profile of the debt tranches and the risk-reward profile of the equity tranche.

In this work, we extend the framework to account for this dependence between

drug development projects. We introduce a single-factor model with a Gaussian copula to model correlations between pipeline drugs (see Section 6.2.3). This approach allows us to generate correlated phase transitions in our simulations, thus evaluating the probability of default and the financial performance of the RBO more accurately.

We use the parameters proposed by Fagnan et al. [190] for a rare disease portfolio (see Table 6.1). In addition, we update the parameters for duration and phase transitions based on the empirical estimates reported by Wong et al. [17] in a recent study using two large pharmaceutical databases to determine the success rates of clinical trials. Compared to the parameters used in Fagnan et al. [190], our recalibrated simulation results in longer clinical development times (0.6 years longer in phase 1 and 1.1 years longer in phase 2), and lower probabilities of success (14 percentage points lower for phase 1 and 9 percentage points lower for phase 2).

In our simulation, we assume a relatively conservative value of 0.20 for pairwise correlation in phase transitions among drug development projects. Although a literature search has not found any estimates of historical correlation between drug development projects, we believe that the correlation between orphan drugs is likely to be weak, given that a large proportion of orphan diseases have monogenic pathologies that act through largely unrelated mechanisms [189, 191]. Furthermore, appropriate portfolio selection protocols can effectively minimize the correlation between assets. By limiting the maximum number of projects that can be acquired per indication group and target family, we can ensure that pipeline drugs in the portfolio are as dissimilar as possible and any risks of failure are largely idiosyncratic in nature. In later sections, we also perform a sensitivity analysis of our results over a range of probabilities of success and pairwise correlation values.

6.2.3 Gaussian Copula

The Gaussian copula is widely used in quantitative finance to model dependence between assets and compute tail risks in credit portfolios. Due to its analytical tractability and simplicity, the model has become the industry standard for pricing collateralized debt obligations (CDOs) [192, 193].

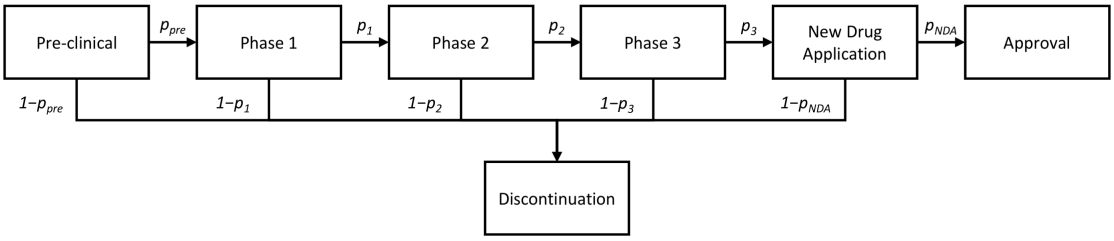


Figure 6-2: Drug development process as a multi-state Markov chain. Each state corresponds to a phase of development. At each phase, pipeline drugs have some probability of advancing to the next higher phase. Drugs that do not successfully advance are discontinued from any further development.

Table 6.1: Parameters used to simulate a megafund with a rare disease portfolio.

	Distribution	**Parameters**	**Pre-clinical**	**Phase 1**	**Phase 2**	**Phase 3**
Upfront cost ($ millions)	Fixed	Constant	3.71			
Milestone cost ($ millions)	Fixed	Constant		3.75	10.00	
Clinical trial cost ($ millions)	Bounded log-normal	Mean	2.80	3.03	7.64	
		Standard deviation	2.29	2.51	7.52	
		Upper bound	10.00	20.00	50.00	
Duration (years)[1]	Log-normal	Mean	2.34	2.14	3.09	
		Standard deviation	1.17	3.79	2.82	
Value ($ millions)	Bounded log-normal	Mean	7.66	24.20	57.80	321.50
		Standard deviation	9.18	28.99	69.24	385.11
		Upper bound	20.00	60.00	200.00	1000.00
		Pairwise correlation	0.20	0.20	0.20	0.20
Phase transition[1]	Bernoulli	Mean	0.80	0.76	0.49	
		Pairwise correlation	0.20	0.20	0.20	0.20

[1] Parameters for phase 1 and phase 2 based on estimates reported by Wong et al. [17].

The copula approach decomposes the problem of estimating multivariate probability distributions into two parts: (1) estimation of the univariate marginal distributions of the underlying random variables, and (2) estimation of the dependence structure between these variables. In particular, the Gaussian copula is defined as:

$$C_\Sigma\Big(\Phi(x_1), \ldots, \Phi(x_n)\Big) = \Phi^n_\Sigma(x_1, \ldots, x_n) \tag{6.1}$$

where C_Σ is the copula function, Φ^n_Σ is the cumulative distribution function of an n-dimensional standard normal distribution with a correlation/covariance matrix Σ that characterizes the dependence structure between the standard normal random variables $\{X_1, \ldots, X_n\}$, and Φ is the cumulative distribution function of a univariate standard normal distribution. A common approach is to assume that dependence is driven by a single common latent factor and write each random variable X_i as the sum of a systematic and an idiosyncratic Gaussian factor. This is known as the single-factor model [194, 195].

$$X_i = \sqrt{\rho_i} \cdot Y + \sqrt{1 - \rho_i} \cdot \epsilon_i \tag{6.2}$$

where $i \in \{1, \ldots, n\}$, $\rho_i \in [0, 1] \, \forall \, i$, $\{Y, \epsilon_1, \ldots, \epsilon_n\}$ are IID standard normal variables, and X_i has a standard normal distribution. In the context of portfolio credit risk modeling, X_i can represent the logarithmic value of each asset in a portfolio; Y can be interpreted as a market risk factor, such as the general state of the economy, that is common to all n assets; and ϵ_i can be interpreted as idiosyncratic risk factors that are specific to each asset. The relative sizes of the systematic and idiosyncratic components are determined by the linear correlation coefficient ρ_i. If $\rho_i = 1 \, \forall \, i$, then all assets have identical values. If $\rho_i = 0 \, \forall \, i$, then the values of all assets are independent of each other. In general, the pairwise correlation between any two assets X_i and X_j, $\text{Corr}(X_i, X_j)$, is given by $\sqrt{\rho_i \rho_j}$. We note that the single-factor model defined in Eq. (6.2) is equivalent to a Gaussian copula model that has a correlation matrix with off-diagonal elements $\Sigma_{ij} = \sqrt{\rho_i \rho_j}$.

The single-factor model can be further simplified by assuming identical pairwise

correlations between all random variables $\{X_1, \ldots, X_n\}$, i.e., $\rho_i = \rho \, \forall \, i$. This reduces the number of parameters in the model from n to just 1. The resulting model is also known in the field of credit risk as the homogeneous large pool Gaussian copula model [192]. The name arises from the assumption of homogeneity in the underlying pool of assets, in terms of identical standard normal marginal distributions and equal correlation coefficients. We make the same assumptions here to generate correlated asset valuations and phase transitions.

We define asset valuations at each stage of development as follows:

$$
\begin{aligned}
B_{ij} &= \min(V_{ij}, M_j) \\
V_{ij} &= \exp(\mu_j + X_{ij} \cdot \sigma_j) \\
X_{ij} &= \sqrt{\rho} \cdot Y + \sqrt{1-\rho} \cdot \epsilon_{ij} \\
\mu_j &= \ln(m_j) - \frac{1}{2} \ln\left(1 + \frac{s_j^2}{m_j^2}\right) \\
\sigma_j &= \sqrt{\ln\left(1 + \frac{s_j^2}{m_j^2}\right)}
\end{aligned}
\tag{6.3}
$$

where V_{ij} is the value of asset i in phase j, m_j and s_j are the estimated mean and standard deviation, respectively, of the asset value distribution at phase j, M_j is the upper bound of possible valuations of assets in phase j, and B_{ij} is the upper-bounded value of asset i in phase j. The parameters used for each stage of development j are summarized in Table 6.1. We draw a single systematic factor Y for each Monte Carlo simulation path.

We define phase transitions at each stage of development in a similar manner:

$$
\begin{aligned}
T_{ij} &= \begin{cases} \text{Success} & \text{if } X_{ij} < \Phi^{-1}(p_j) \\ \text{Failure} & \text{otherwise} \end{cases} \\
X_{ij} &= \sqrt{\rho} \cdot Y + \sqrt{1-\rho} \cdot \epsilon_{ij}
\end{aligned}
\tag{6.4}
$$

where p_j is the unconditional probability of success for phase j, Φ^{-1} is the inverse cumulative distribution function of a univariate standard normal distribution, T_{ij} is

the outcome of drug development project i in phase j (i.e., whether the project has successfully advanced to the next higher phase $(j+1)$). We draw different sets of Y and ϵ_{ij} for asset valuations and for phase transitions. The parameters used for each stage of development j are summarized in Table 6.1.

While we have focused on a flat correlation structure in the derivations and in our analysis, the Gaussian copula single-factor model can be easily extended to arbitrary correlation structures by using multi-factor models. In addition, it is straightforward to generalize the single-factor model for non-Gaussian distributions with zero mean and unit variance.

6.2.4 Impact of Correlation on Tail Risk

Suppose there are n correlated projects in a portfolio, each with an unconditional probability of success equal to p. We can derive the distribution function of the number of successes in the portfolio by rewriting Eq. (6.4). Using the property of conditional independence, we observe that, given a realization of the systematic factor, successes are independent with probability:

$$\begin{aligned} P[T = \text{Success} \mid Y = y] &= \Phi\left(\frac{\Phi^{-1}(p) - \sqrt{\rho} \cdot y}{\sqrt{1-\rho}}\right) \\ &= p_y \end{aligned} \tag{6.5}$$

The probability of having exactly m successes out of n projects is thus:

$$\begin{aligned} P[H = m] &= \int_{-\infty}^{\infty} P[H = m \mid Y = y] \cdot \phi(y)\, \mathrm{d}y \\ &= \int_{-\infty}^{\infty} \binom{n}{m} \cdot {p_y}^m \cdot (1-p_y)^{(n-m)} \cdot \phi(y)\, \mathrm{d}y \end{aligned} \tag{6.6}$$

where ϕ is the probability density function of the standard normal distribution. We

can approximate the integral in Eq. (6.6) using the Gauss-Hermite quadrature:

$$P[H = m] \approx \frac{1}{\sqrt{\pi}} \sum_{i=1}^{k} w_i \cdot \binom{n}{m} \cdot {p_{x_i}}^m \cdot (1 - p_{x_i})^{(n-m)} \tag{6.7}$$
$$p_{x_i} = \Phi\left(\frac{\Phi^{-1}(p) - \sqrt{2\rho} \cdot x_i}{\sqrt{1-\rho}}\right)$$

where k is the number of nodes used, and $\{x_i\}_{i=1}^k$ and $\{w_i\}_{i=1}^k$ are the corresponding abscissas and weights, respectively. To illustrate the impact of correlation on tail risk, we compute for homogeneous portfolios of different sizes, unconditional probabilities of success, and pairwise correlations—$n = \{20, 50, 100\}$, $p = \{0.1, 0.3, 0.5\}$, and $\rho = \{0.0, 0.10, 0.2, 0.3, 0.4\}$, respectively—their distribution functions of successes. Table 6.2 and Fig. 6-3 summarize the results.

From Table 6.2, we find that correlation has no effect on the expected number of successful projects in the portfolio. However, it has significant impact on the variability of the distribution of successes. For a portfolio with $n = 20$ projects and $p = 0.1$, a pairwise correlation of $\rho = 0.1$ increases the standard-deviation-to-mean ratio (a standardized measure of dispersion of the probability distribution; also known as the coefficient of variation [CV]) of the distribution of successes by 31.3% relative to the case with zero correlation. In addition, we observe that the increase in CV is greater with higher unconditional probabilities of success, *ceteris paribus*: For a portfolio with $n = 20$ and $p = 0.5$, a pairwise correlation of $\rho = 0.1$ increases the CV by 50.0% relative to a portfolio with the same parameters but zero correlation. The effect is also more pronounced as the size of the portfolio becomes larger: For $n = 100$ and $p = 0.1$, the increase is 116.7%.

In general, the impact of correlation seems to be weaker for portfolios with smaller n and p, likely because the distributions of successes under such parameters are very skewed to begin with (see Fig. 6-3). Nevertheless, it is clear that positive correlation between assets reduces benefits from diversification by introducing systematic risk to the portfolio. This has major implications for the megafund because returns are directly related to portfolio performance (i.e., the number of successful projects).

Table 6.2: Distribution functions of successes computed using Gauss-Hermite quadrature with $k = 30$ nodes.

	$n = 20$			$n = 50$			$n = 100$		
ρ	μ	σ	σ/μ	μ	σ	σ/μ	μ	σ	σ/μ
				$\rho = 0.1$					
0.0	2	1.34	0.67	5	2.12	0.42	10	3.00	0.30
0.1	2	1.75	0.88	5	3.56	0.71	10	6.48	0.65
0.2	2	2.13	1.06	5	4.70	0.94	10	8.96	0.90
0.3	2	2.49	1.25	5	5.74	1.15	10	11.14	1.11
0.4	2	2.85	1.43	5	6.73	1.35	10	13.19	1.32
				$\rho = 0.3$					
0.0	6	2.05	0.34	15	3.24	0.22	30	4.58	0.15
0.1	6	2.98	0.50	15	6.37	0.42	30	11.93	0.40
0.2	6	3.70	0.62	15	8.46	0.56	30	16.37	0.55
0.3	6	4.32	0.72	15	10.20	0.68	30	19.97	0.67
0.4	6	4.90	0.82	15	11.74	0.78	30	23.15	0.77
				$\rho = 0.5$					
0.0	10	2.24	0.22	25	3.54	0.14	50	5.00	0.10
0.1	10	3.33	0.33	25	7.18	0.29	50	13.52	0.27
0.2	10	4.14	0.41	25	9.54	0.38	50	18.50	0.37
0.3	10	4.84	0.48	25	11.46	0.46	50	22.47	0.45
0.4	10	5.47	0.55	25	13.15	0.53	50	25.95	0.52

In reality, drug development programs tend to exhibit some extent of correlation with one another depending on the similarities in their underlying scientific pathways, mechanisms and targets. Therefore, the risk of the megafund portfolio will be underestimated if we assume that the projects are independent. To obtain accurate estimates of default probabilities and investment performance, it is critical that model used for phase transitions incorporates some form of dependence between pipeline drugs.

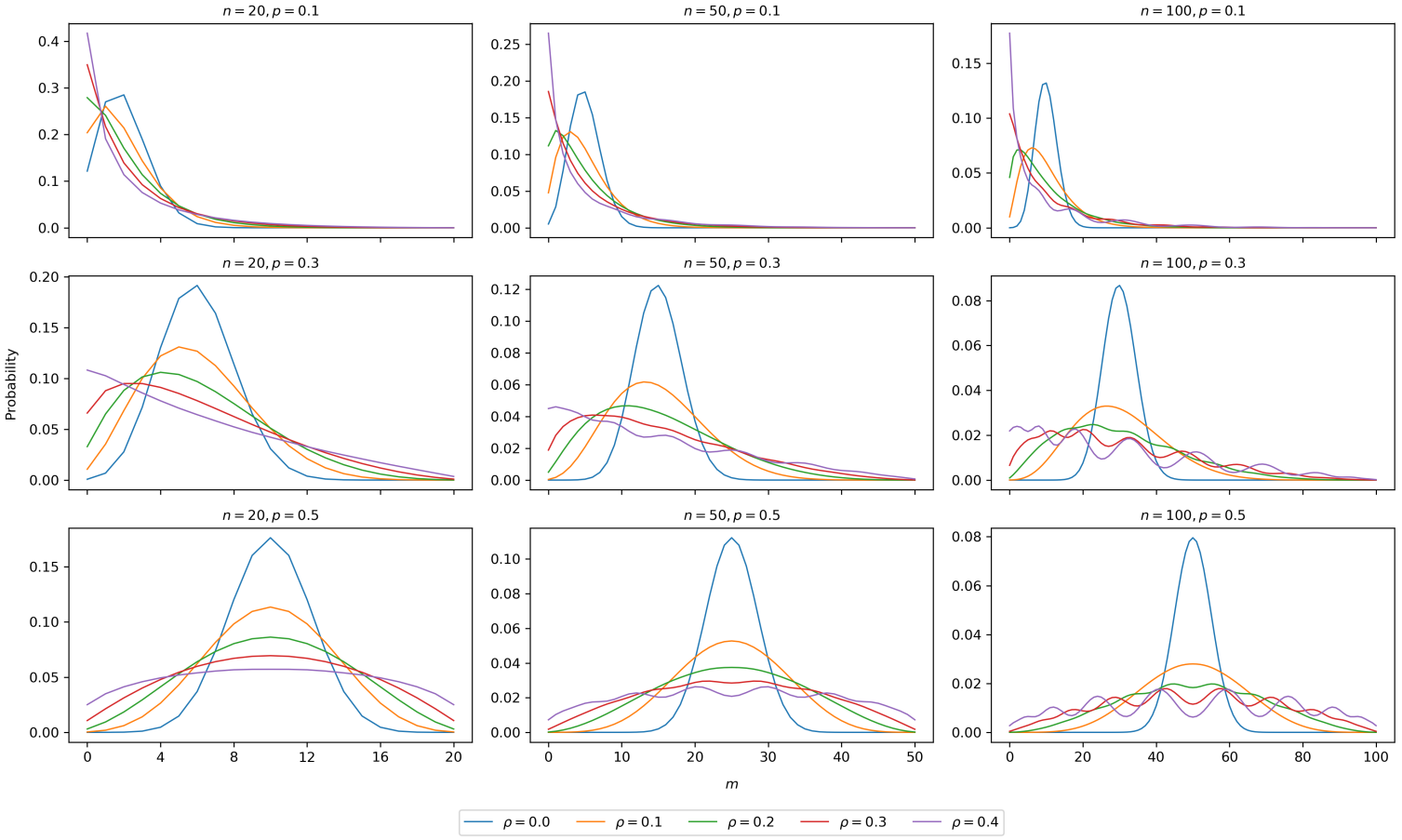


Figure 6-3: Distribution functions of successes computed using Gauss-Hermite quadrature with $k = 30$ nodes. We find that the impact of correlation on the distribution of the number of successes depends on both the number of projects and their unconditional probabilities of success.

6.3 Results

6.3.1 Simulation

We perform three sets of experimental simulations of an orphan drug megafund. In the first, we simulate the performance of a "vanilla" megafund as outlined in Section 6.2. In the second, we consider an RBO structure identical to the first, except that it incorporates an additional credit enhancement mechanism proposed by Fagnan et al. [196]. We assume that a third party is willing to take on some of the downside risk to debtholders by providing a guarantee for part of the debt issued, up to a maximum value of $100M. This funding guarantee serves as a form of collateral that can be used to make up any shortfall in cash flow to meet debt obligations during the tenor of the fund. This type of external credit support may be provided to the megafund by a government agency, a private foundation, or even a patient advocacy group to advance a scientific or medical cause (e.g., drug development for a specific rare disease).

In our third experiment, we consider for comparative purposes an all-equity financing structure, while keeping all other modeling assumptions unchanged, to demonstrate the advantages of leverage and diversification. We assume that this megafund begins with an initial amount of capital of $275M, the size of the equity tranche in the first two sets of experiments. With a smaller pool of investable capital, the equity-only fund can only afford to acquire and finance 11 pre-clinical compounds for its portfolio, as opposed to 23 in the other two experiments.

For each experiment, we perform 2,000,000 Monte Carlo simulated paths of drug development, drawing from the random distributions parameterized in Section 6.2 for each realization. By aggregating the results for each RBO structure—vanilla, guarantee-backed, and equity-only—we compute the risk profile of the debt tranches, the distribution of returns of the equity tranche, the expected cost of guarantee, and the impact of the research, quantified by the number of compounds sold in phases 2 and 3. The results are summarized in Fig. 6-4 and Table 6.3.

We find that the risk of bond default is very small for both the senior and junior

debt tranches under the vanilla megafund structure. The probability of default for the senior tranche is less than 1 basis point (bp), comparable to the historical default rate of AAA-rated corporate bonds. The default rate of the mezzanine tranche is higher at 55 bp, but still well below the average default rate of investment-grade corporate bonds over the same time horizon (see Appendix E.1). With the addition of a third-party funding guarantee, the default rates of both tranches fall to zero. This effectively makes the junior tranche a second senior tranche. We therefore combine both debt tranches in the guarantee-backed megafund into a single senior debt issue in our treatment. Despite the high face value of the guarantee, we note that the expected cost to the guarantor is actually very small, about $37,000.

The vanilla megafund outperforms the all-equity financing structure in equity returns. It achieves an expected annualized return on equity (ROE) of 11.0%, 2.8 percentage points higher than that of the equity-only fund. Moreover, the probability of substantial gains, defined as an annualized ROE in excess of 25%, is four times higher under the standard structure (14.4%) that the equity-only fund (3.6%). Its Sharpe ratio, however, is about 3 percentage points lower in comparison.

Although the chances of a wipeout in the leveraged megafund are slightly higher than in the all-equity structure (0.6% versus 0.0%), the probability of a loss to equity is lower overall (19.4% versus 24.5%). In general, we find that distribution of cumulative ROE has a fatter left tail under the all-equity structure than under the vanilla structure (see Fig. 6-4), suggesting that the use of leverage helps to reduce the downside risk and improve the upside potential.

With the addition of a funding guarantee, we observe a modest improvement in the return profile. The probability of loss falls further to 18.7%, while the expected annualized ROE improves slightly to 11.5%. The Sharpe ratio for the guarantee-backed structure is also the highest among the three RBO structures (66.1%), suggesting that the presence of a guarantee can help to reduce volatility without compromising returns. (As a reference point, the average return and the corresponding Sharpe ratio of the Center for Research in Security Prices (CRSP) value-weighted index between 1970 and 2016 were 10.9% and 37%, respectively [25].)

Among the three RBO structures, the leveraged structure performs the best in terms of research impact. On average, 9.5 out of 23 pre-clinical projects in the vanilla megafund portfolio reach either phase 2 or phase 3 by the end of the simulation horizon, the rest discontinued or sold at earlier phases. In contrast, the equity-only fund starts with 11 investigational compounds in its portfolio, out of which typically only 4.1 are successfully liquidated at either phase 2 or phase 3, less than half that of its leveraged counterpart.

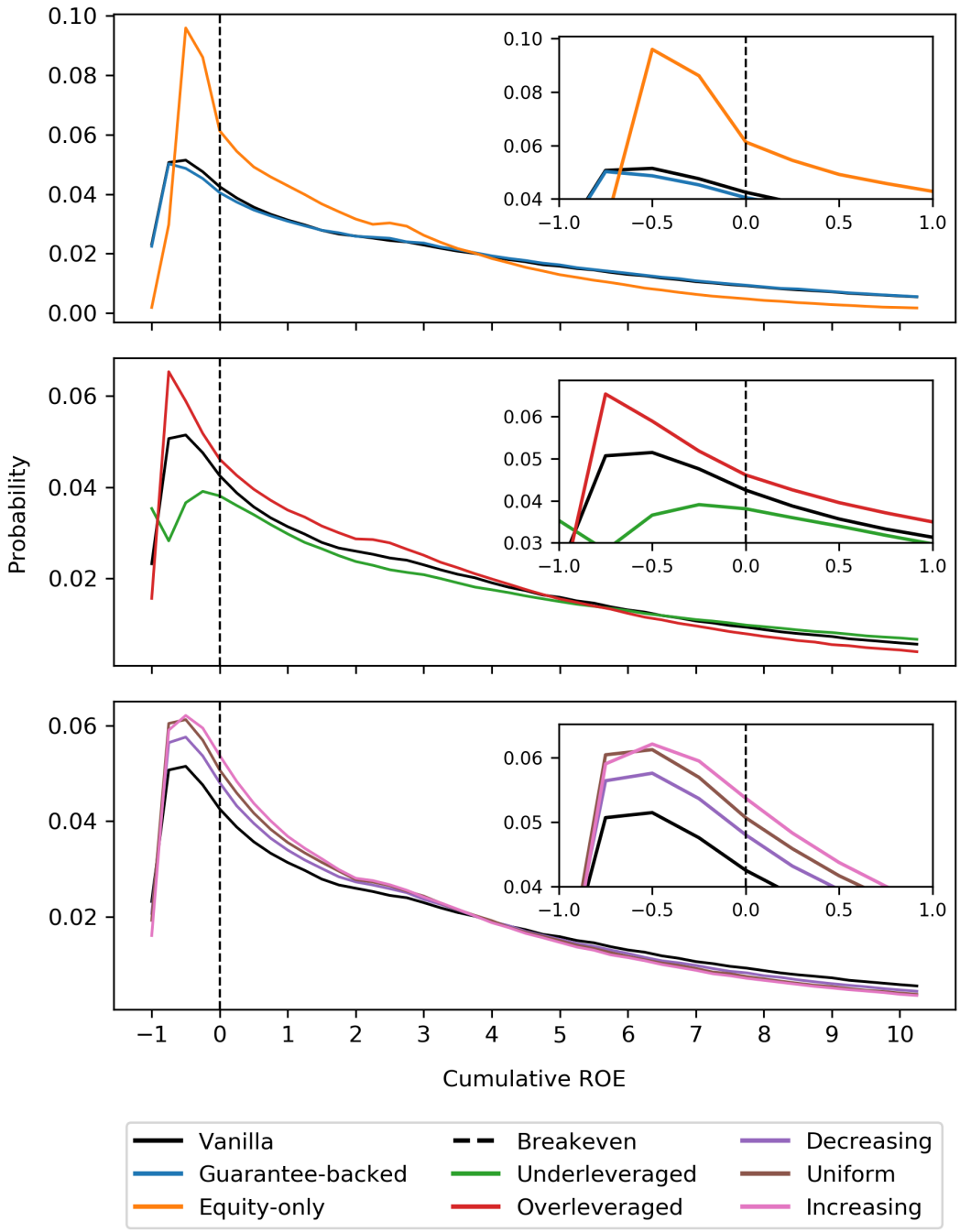


Figure 6-4: Distributions of cumulative ROE for different RBO structures (top), capital structures (middle), and acquisition strategies (bottom). We truncate the distributions when cumulative ROE equals 10x for better visualization. The distributions demonstrate a positive skew with higher-than-normal kurtosis (leptokurtic).

Table 6.3: Performance of each RBO structure over 2,000,000 Monte Carlo simulation paths. *Abbreviations*: ann, annualized.

	RBO Structure							
	Vanilla	**Guarantee-backed**	**Equity-only**	**Under-leveraged**	**Over-leveraged**	**Decreasing**	**Uniform**	**Increasing**
Structure								
Capital								
Total ($ millions)	575	575	275	425	875	575	575	575
Senior tranche ($ millions)	250	300		125	500	250	250	250
Junior tranche ($ millions)	50			25	100	50	50	50
Equity tranche ($ millions)	275	275	275	275	275	275	275	275
Guarantee								
Total ($ millions)		100						
Portfolio								
Number acquired at pre-clinical	23	23	11	17	35	23	23	23
Performance								
Senior tranche								
Prob of default (bp)	0.8	0.0		0.0	26.9	0.1	0.0	0.0
Expected loss (bp)	0.0	0.0		0.0	1.7	0.0	0.0	0.0
Junior tranche								
Prob of default (bp)	54.8			4.0	266.1	30.9	21.5	13.7
Expected loss (bp)	16.1			0.7	143.7	7.3	4.6	2.6
Equity tranche								
Expected cumulative ROE	3.8	3.8	2.1	3.0	5.0	3.2	2.9	2.8
Cumulative ROE sd	4.5	4.5	2.7	3.6	5.9	4.0	3.7	3.5
Cumulative ROE skewness	1.6	1.6	1.6	1.5	1.8	1.6	1.6	1.6
Cumulative ROE kurtosis	3.3	3.2	3.1	3.1	4.1	3.4	3.4	3.4
Expected ann ROE (%)	11.0	11.5	8.2	9.8	11.5	9.9	9.3	9.2
Sharpe ratio[1] (%)	58.2	66.1	61.2	62.6	42.6	56.6	54.7	56.6
Prob of ann ROE $= -1.00$ (%)	0.6	0.3	0.0	0.1	2.7	0.3	0.2	0.2
Prob of ann ROE < 0.00 (%)	19.4	18.7	24.5	21.5	15.8	21.3	22.4	22.4
Prob of ann ROE > 0.10 (%)	59.2	60.4	46.1	54.7	64.1	55.3	53.0	51.9
Prob of ann ROE > 0.25 (%)	14.4	14.7	3.6	8.9	21.7	10.8	8.9	8.2
Guarantee								
Prob of draw (%)		0.3						
Expected cost[2] ($ thousands)		37.4						
Research impact								
Number sold in phase 2	5.3	5.1	1.5	3.2	10.2	5.0	4.6	4.9
Number sold in phase 3	4.2	4.3	2.6	3.6	4.9	3.7	3.4	3.2

[1] Risk-free rate 2.0%. [2] Net present value at 2.0% discount rate.

6.3.2 Sensitivity Analysis

We perform a sensitivity analysis of our results with respect to several key parameters in our framework, namely the capital structure, the acquisition strategy, and the correlation and probability of success at phase transition.

Capital Structure

In the previous section, we assumed a relatively well-balanced capital structure with a debt-to-equity ratio of 1.09 for the vanilla megafund. To examine the impact of different capital structures on performance, we consider two additional configurations. The first assumes an underleveraged capital structure with the same amount of equity as the vanilla case ($275M), but half as much debt ($150M). The second assumes an overleveraged structure with also the same amount of equity ($275M), but twice as much debt ($300M). The resulting debt-to-equity ratios for the underleveraged megafund and the overleveraged megafund are 0.55 and 2.18, respectively. We summarize their performance in Fig. 6-4 and Table 6.3.

We find that the risk of bond default generally increases with the leverage ratio of the capital structure. In the underleveraged megafund, the equity tranche (the tranche that absorbs the first loss to capital) is almost twice as large as the debt tranches combined. This high level of overcollateralization allows the fund to remain solvent over a wide range of portfolio losses. Assuming a zero-coupon bond, the underleveraged megafund can lose up to 62% of its portfolio and still have enough capital to repay all of its debt obligations.

In contrast, the size of the equity tranche in the overleveraged megafund is less than half that of the debt tranches. Consequently, a small shock to the portfolio can easily wipe out the entire equity tranche and force the megafund into default. Assuming a zero-coupon bond, the overleveraged megafund must not lose more than 26% of its portfolio in order to have sufficient funds to redeem its bonds. The probabilities of default are therefore much larger for the overleveraged capital structure than for the balanced and the underleveraged structures.

By issuing more debt, the overleveraged megafund can acquire and finance a larger number of projects (35 versus 17 for the underleveraged megafund). With a larger and therefore more diversified portfolio, its expected ROE is correspondingly higher. Its Sharpe ratio, however, is the lowest among the three capital structures, suggesting that the improvement in returns is outweighed by the increase in volatility associated with the use of greater leverage. We observe the opposite for the underleveraged megafund, which has the lowest expected ROE but the highest Sharpe ratio.

The megafund demonstrates very different risk-reward characteristics under each of these capital structures. In general, the use of leverage helps to improve the performance of the megafund. However, it comes at the cost of increased risk to bondholders and also greater volatility in returns. The capital structure of a megafund should therefore be carefully selected to maximize the ROE while keeping the Sharpe ratio and the default rates attractive to equity holders and fixed-income investors. To avoid under-borrowing and over-borrowing, the leverage ratio should be optimized based on the cost, value, and risk profiles of the underlying assets in the portfolio.

Acquisition Strategy

In the next step of our sensitivity analysis, instead of assuming that all assets are acquired at the start of the simulation, we consider an alternative strategy in which a small number of projects is acquired each period until the target capacity is reached (i.e., the portfolio is built up over time). Under some conditions, this strategy may be a more realistic example of a potential business model for an orphan drug megafund. The earlier assumption is useful if there is a large pool of projects that is readily available for immediate investment, e.g., the rare diseases therapeutic development program at the National Center for Advancing Translational Sciences [190]. In other cases, there may not be enough projects of sufficient quality on the market to create a strong and well-diversified portfolio.

Instead of settling for mediocre opportunities, a strategy of rolling acquisitions gives portfolio managers more time to source, evaluate, and identify promising clinical assets for acquisition. This approach leaves room for potential investment in

breakthroughs that may emerge after the inception of the fund. Moreover, it aligns with the typical operation of translational drug development grant programs, which screen a large number of proposals annually, while enrolling only a few high-potential projects that have innovative scientific approaches or target unmet clinical needs.

Here, we consider three different acquisition patterns. We assume that the vanilla megafund either makes a monotonically increasing number of acquisitions each period, a uniform number, or a monotonically decreasing number, until the portfolio contains 23 projects. We find that the expected annualized ROE and research impact are smaller under rolling acquisitions than under our original assumption (see Table 6.3). This is not surprising, since each stage of drug development requires a certain amount of time for clinical testing. Under a rolling acquisition strategy, a part of the portfolio is acquired after the first period. These projects are generally financed and developed for a shorter duration than those acquired at the beginning. As a result, they are less likely to complete phase 2 within the time horizon of the simulation before the portfolio must be liquidated. The probability of default in the junior tranche is consequently lower, because fewer risky late-stage drug development programs need to be funded. (The probability of transition is the lowest for phase 2 to phase 3.) As a trade-off, the expected ROE is also smaller because more drugs are sold before they can reach phase 3, which has the highest sale value. The effect is the greatest under the monotonically increasing pattern, in which the largest part of the portfolio is acquired later in the simulation.

Correlation and Probability of Success

Finally, we investigate the sensitivity of our results to different pairwise correlations in phase transitions and the probabilities of success. We vary the correlation between 0% and 40%, and adjust the probabilities of success for all phases by -10%, 0%, and +10%. For each combination of RBO structure, correlation value, and adjustment to the probability of success, we perform 100,000 Monte Carlo simulation paths. We summarize the results in Tables 6.4 to 6.6 and Fig. 6-5.

Intuitively, the expected number of projects that reach phase 3 increases with

the probability of success. We also observe a corresponding increase in expected returns with higher adjusted phase transition probabilities, since the sale value of assets is substantially higher for late-stage projects than for early-stage drugs in the pipeline. As shown in Tables 6.4 to 6.6, a relative adjustment of +10% to the baseline probability of success at each stage of development improves the expected annualized ROE of the vanilla megafund by 4 percentage points, while the same adjustment in the opposite direction reduces the ROE by about 4.3 percentage points. We observe similar trends for the guaranteed-backed and the equity-only megafunds.

In Fig. 6-5, we plot the distribution of cumulative ROE for the different RBO structures, correlations, and adjustments to the probabilities of success. Because correlated projects tend to have similar outcomes, we find that the risk of tail events generally increases with the correlation between projects in the portfolio. This can be seen from the large positive skew and the heavy tails that highly correlated portfolios show in their distributions. We observe improvements in the expected cumulative ROE when the correlation is increased, but this is likely the effect of outliers in the right tail—that is, rare events where a large number of correlated projects reach phase 3 simultaneously, thus giving rise to extremely high returns. The mean of the annualized ROE is less sensitive to these outliers (see Fig. 6-6). In fact, the Sharpe ratio demonstrates an inverse relationship to the correlation (see Tables 6.4 to 6.6), indicating that greater correlation actually leads to lower annualized returns and greater volatility.

In most cases, the vanilla megafund outperforms the equity-only structure, except in the worst-case scenario, where the probabilities of success are low and the correlation between projects is high. The expected number of successful projects is small under this set of parameters, and it is thus unlikely that the megafund can generate sufficient cash flow to sustain its debt obligations and investment activities under these conditions. The high level of correlation further exacerbates the situation by introducing substantial systematic risk to the portfolio. It is clear that, given the risk profile of the underlying portfolio, the megafund is overleveraged. In such cases, a better performing megafund could be created by either adopting a more

appropriate capital structure or securing some form of funding guarantee, i.e., the guarantee-backed structure.

Despite the variation in parameter values, the probability of default for the senior tranche remains below 1 bp in almost all scenarios. This can be attributed to the credit enhancement mechanisms adopted in the RBO structure, including the subordination of cash flows and the interest coverage tests to trigger early liquidation during periods of illiquidity. The risk of default for the junior tranche, however, is very sensitive to changes in either parameter. Like the trends observed in equity returns, the risk of default increases with the level of correlation between projects, when there is a greater probability of loss, but decreases with the probabilities of success, when there is a greater probability of profit. With a funding guarantee in place, the probability of default for the guarantee-backed megafund is consistently kept below 0.1 bp. The expected cost to the guarantee also increases with the level of correlation and decreases with the probabilities of success.

In general, we find that the performance of the megafund becomes less attractive when correlation between projects is introduced. Nevertheless, the vanilla megafund outperforms the all-equity structure over a wide range of probabilities of success and correlation, except in cases where there is substantial deviation from the presumed values. In those scenarios, the capital structure and leverage ratio need to be re-optimized with respect to the risk profile of the underlying portfolio. The use of a funding guarantee can also greatly improve the performance of the megafund. Overall, the risk of default for the senior tranche remains close to zero even when large correlations and small probabilities of success are assumed.

Table 6.4: Sensitivity of the vanilla RBO performance to different pairwise correlations between phase transitions and probabilities of success. The results are based on 100,000 Monte Carlo simulation paths for each combination of pairwise correlation and probability of success. *Abbreviations*: ρ, pairwise correlation between phase transitions; p, probabilities of success for pre-clinical, phase 1, and phase 2; ann, annualized.

	0.9p					**1.0p**					**1.1p**				
ρ	**0.0**	**0.1**	**0.2**	**0.3**	**0.4**	**0.0**	**0.1**	**0.2**	**0.3**	**0.4**	**0.0**	**0.1**	**0.2**	**0.3**	**0.4**
Senior tranche															
Prob of default (bp)	1.2	0.9	0.9	1.1	0.9	1.0	0.5	0.5	0.4	0.4	0.2	0.5	0.4	0.6	0.5
Expected loss (bp)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Junior tranche															
Prob of default (bp)	41.1	71.1	99.6	118.2	131.8	20.3	36.0	54.1	75.3	87.3	9.9	16.6	26.5	37.5	49.8
Expected loss (bp)	15.1	20.7	24.7	27.6	28.8	8.6	12.5	16.0	19.3	20.7	4.6	6.7	9.1	11.8	13.6
Equity tranche															
Expected cumulative ROE	2.0	2.4	2.7	3.1	3.4	3.1	3.4	3.8	4.1	4.4	4.4	4.7	5.0	5.3	5.5
Cumulative ROE sd	2.5	3.3	4.0	4.6	5.1	3.0	3.8	4.5	5.0	5.5	3.6	4.4	5.0	5.5	5.9
Cumulative ROE skewness	1.5	1.7	1.8	1.8	1.8	1.3	1.5	1.6	1.6	1.5	1.1	1.4	1.4	1.4	1.3
Cumulative ROE kurtosis	2.8	4.2	4.5	4.3	3.7	2.1	3.2	3.3	3.0	2.6	1.7	2.4	2.4	2.2	1.9
Expected ann ROE (%)	7.8	7.2	6.7	6.3	6.2	11.9	11.4	11.0	10.6	10.3	15.7	15.3	15.0	14.7	14.3
Sharpe ratio[1] (%)	46.8	33.6	26.1	22.1	20.0	92.9	71.4	58.5	49.2	44.3	142.1	117.4	99.0	86.0	76.0
Prob of ann ROE $= -1.00$ (%)	0.4	0.7	1.1	1.3	1.4	0.2	0.4	0.6	0.8	1.0	0.1	0.2	0.3	0.4	0.5
Prob of ann ROE < 0.00 (%)	20.9	26.1	29.2	31.5	33.3	10.9	15.9	19.3	22.2	24.4	5.3	8.3	11.3	13.8	16.1
Prob of ann ROE > 0.10 (%)	46.4	47.3	48.0	48.6	49.2	62.2	60.3	59.3	58.8	58.5	76.0	72.6	70.4	69.0	67.9
Prob of ann ROE > 0.25 (%)	2.6	6.2	9.4	12.4	15.2	6.7	10.9	14.5	17.5	20.2	14.2	18.2	21.3	23.8	26.0
Research impact															
Number sold in phase 2	4.3	4.4	4.5	4.5	4.6	5.2	5.3	5.3	5.4	5.4	6.2	6.2	6.3	6.3	6.3
Number sold in phase 3	2.5	2.9	3.2	3.6	3.9	3.6	3.9	4.2	4.6	4.9	4.8	5.1	5.4	5.7	5.9

[1] Risk-free rate 2.0%.

Table 6.5: Sensitivity of the guarantee-backed RBO performance to different pairwise correlations between phase transitions and probabilities of success. The results are based on 100,000 Monte Carlo simulation paths for each combination of pairwise correlation and probability of success. *Abbreviations*: ρ, pairwise correlation between phase transitions; p, probabilities of success for pre-clinical, phase 1, and phase 2; ann, annualized.

	0.9p					**1.0p**					**1.1p**				
ρ	**0.0**	**0.1**	**0.2**	**0.3**	**0.4**	**0.0**	**0.1**	**0.2**	**0.3**	**0.4**	**0.0**	**0.1**	**0.2**	**0.3**	**0.4**
Senior tranche															
Prob of default (bp)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Expected loss (bp)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Equity tranche															
Expected cumulative ROE	2.1	2.4	2.8	3.1	3.5	3.2	3.5	3.9	4.2	4.5	4.5	4.8	5.1	5.3	5.6
Cumulative ROE sd	2.5	3.3	4.0	4.6	5.1	3.0	3.8	4.5	5.1	5.6	3.7	4.4	5.0	5.5	5.9
Cumulative ROE skewness	1.4	1.7	1.8	1.8	1.7	1.2	1.5	1.6	1.6	1.5	1.1	1.3	1.4	1.3	1.3
Cumulative ROE kurtosis	2.6	4.2	4.6	4.2	3.6	1.9	3.2	3.2	2.9	2.6	1.6	2.4	2.3	2.1	1.9
Expected ann ROE (%)	8.2	7.7	7.5	7.3	7.3	12.3	11.8	11.5	11.2	11.1	16.0	15.6	15.3	15.1	14.8
Sharpe ratio[1] (%)	54.0	41.1	35.1	31.1	29.3	99.1	78.5	66.0	58.7	54.1	147.5	122.8	105.8	94.2	84.9
Prob of ann ROE $= -1.00$ (%)	0.3	0.4	0.4	0.5	0.5	0.2	0.2	0.3	0.4	0.4	0.1	0.1	0.2	0.2	0.2
Prob of ann ROE < 0.00 (%)	19.7	25.2	28.4	30.9	32.7	10.2	15.0	18.6	21.5	23.9	4.8	7.8	10.7	13.4	15.8
Prob of ann ROE > 0.10 (%)	47.7	48.3	49.1	49.6	50.2	63.6	61.5	60.4	59.7	59.4	77.5	73.8	71.7	70.1	68.9
Prob of ann ROE > 0.25 (%)	2.6	6.3	9.6	12.6	15.5	6.8	11.2	14.7	17.8	20.5	14.7	18.6	21.7	24.3	26.7
Guarantee															
Prob of draw (%)	0.3	0.4	0.4	0.5	0.5	0.2	0.2	0.3	0.4	0.4	0.1	0.1	0.2	0.2	0.2
Expected cost[2] ($ thousands)	33.7	38.6	41.4	44.1	45.3	26.8	31.8	35.0	37.0	35.5	18.8	21.9	25.8	27.2	29.0
Research impact															
Number sold in phase 2	4.2	4.2	4.3	4.4	4.4	5.1	5.1	5.1	5.2	5.2	6.0	6.0	6.1	6.1	6.1
Number sold in phase 3	2.6	2.9	3.3	3.6	4.0	3.6	4.0	4.3	4.6	4.9	4.9	5.2	5.5	5.7	6.0

[1] Risk-free rate 2.0%. [2] Net present value at 2.0% discount rate.

Table 6.6: Sensitivity of the equity-only RBO performance to different pairwise correlations between phase transitions and probabilities of success. The results are based on 100,000 Monte Carlo simulation paths for each combination of pairwise correlation and probability of success. *Abbreviations*: ρ, pairwise correlation between phase transitions; p, probabilities of success for pre-clinical, phase 1, and phase 2; ann, annualized.

	0.9p					**1.0p**					**1.1p**				
ρ	**0.0**	**0.1**	**0.2**	**0.3**	**0.4**	**0.0**	**0.1**	**0.2**	**0.3**	**0.4**	**0.0**	**0.1**	**0.2**	**0.3**	**0.4**
Equity tranche															
Expected cumulative ROE	1.2	1.4	1.6	1.7	1.9	1.8	1.9	2.1	2.3	2.4	2.5	2.6	2.8	2.9	3.0
Cumulative ROE sd	1.7	2.1	2.4	2.7	2.9	2.0	2.4	2.7	3.0	3.2	2.4	2.7	3.0	3.2	3.4
Cumulative ROE skewness	1.5	1.7	1.8	1.8	1.8	1.4	1.5	1.6	1.6	1.5	1.2	1.3	1.4	1.3	1.3
Cumulative ROE kurtosis	2.9	4.1	4.4	4.3	3.9	2.4	3.0	3.2	3.1	2.7	1.9	2.3	2.3	2.1	1.9
Expected ann ROE (%)	5.2	5.5	5.9	6.2	6.6	8.0	8.0	8.2	8.4	8.6	10.8	10.7	10.7	10.7	10.7
Sharpe ratio[1] (%)	37.1	37.8	39.4	40.8	42.3	69.6	63.8	61.3	59.8	59.4	106.8	93.7	86.4	81.6	78.4
Prob of ann ROE $= -1.00$ (%)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Prob of ann ROE < 0.00 (%)	29.1	31.6	33.6	35.2	36.6	18.4	21.8	24.4	26.6	28.4	10.4	13.7	16.4	18.8	20.8
Prob of ann ROE > 0.10 (%)	31.1	34.1	36.4	38.2	40.0	43.2	44.9	46.1	47.3	48.4	56.6	56.4	56.6	56.8	57.1
Prob of ann ROE > 0.25 (%)	0.3	1.2	2.2	3.3	4.4	1.0	2.3	3.6	4.9	6.2	2.6	4.2	5.7	7.1	8.4
Research impact															
Number sold in phase 2	1.2	1.2	1.2	1.3	1.3	1.4	1.5	1.5	1.5	1.6	1.8	1.8	1.8	1.8	1.9
Number sold in phase 3	1.7	1.9	2.0	2.2	2.4	2.3	2.5	2.6	2.8	3.0	3.0	3.2	3.3	3.4	3.6

[1] Risk-free rate 2.0%.

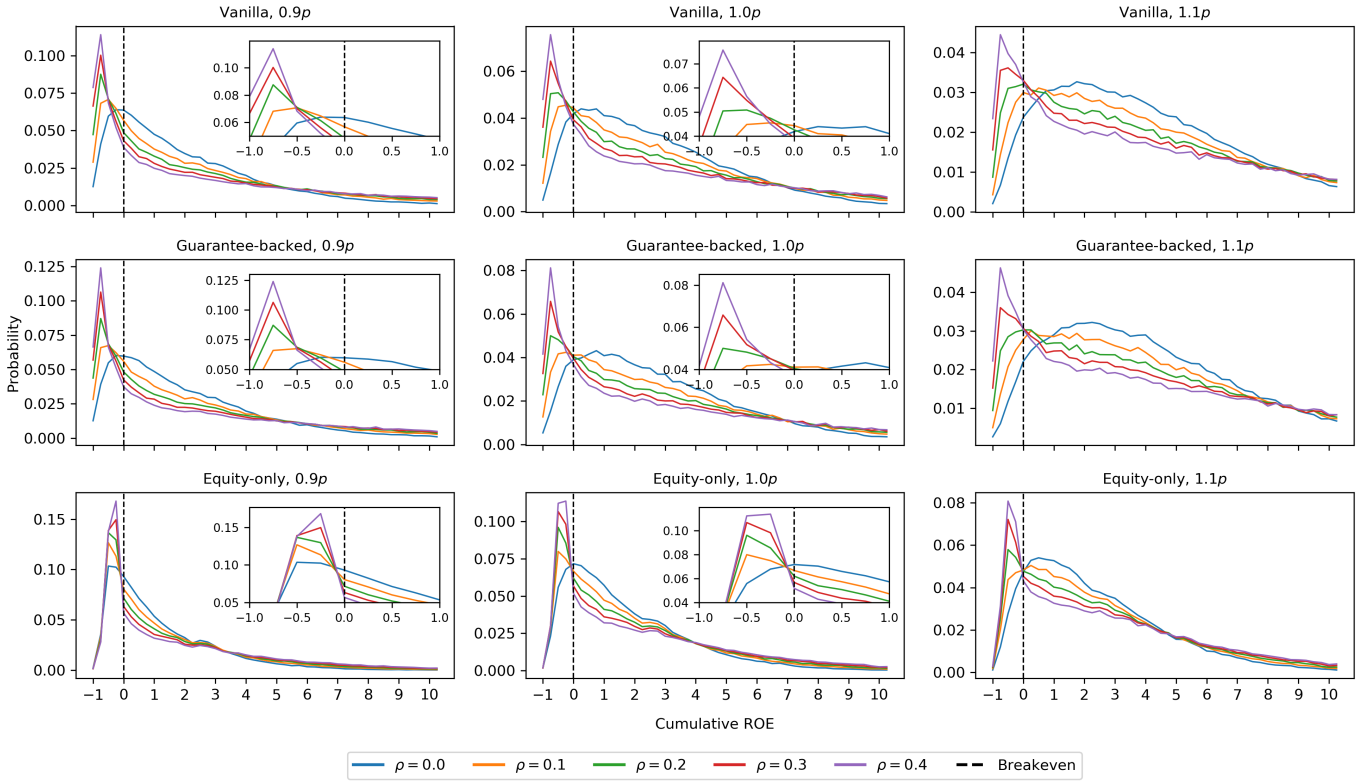


Figure 6-5: Sensitivity of cumulative ROE to different pairwise correlations between phase transition and probabilities of success. Each plot corresponds to a different combination of RBO structure and adjustment to the probability of success. We truncate the plots when cumulative ROE equals 10x for better visualization. *Abbreviations*: ρ, pairwise correlation between phase transitions; p, probabilities of success for pre-clinical, phase 1, and phase 2.

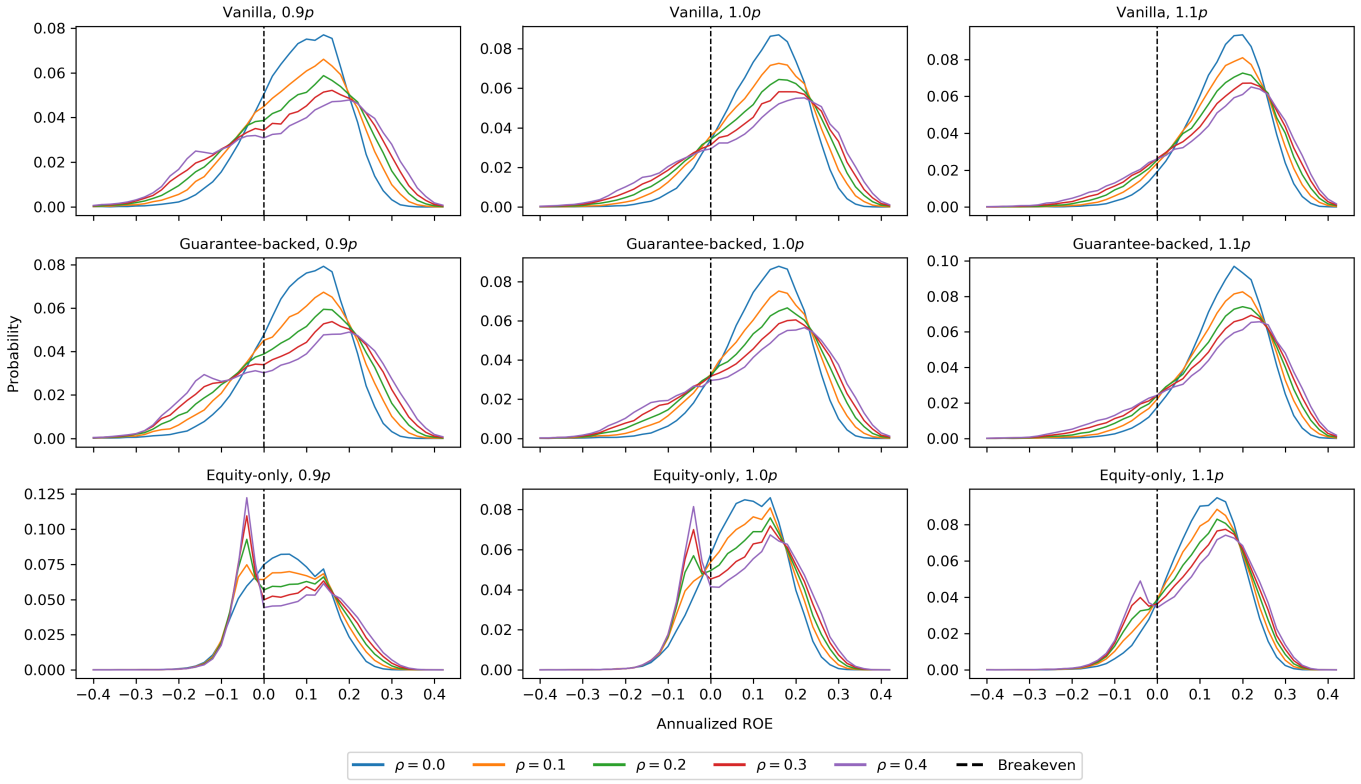


Figure 6-6: Sensitivity of annualized ROE to different pairwise correlations between phase transitions and probabilities of success. Each plot corresponds to a different combination of RBO structure and probability-of-success adjustment. We truncate the plots at annualized ROE -0.4 and +0.4 for better visualization. *Abbreviations*: ρ, pairwise correlation between phase transitions; p, probabilities of success for pre-clinical, phase 1, and phase 2.

6.4 Discussion

Traditional financing models have generally struggled to support early-stage drug development, which corresponds to the riskiest and most challenging part of the drug approval process. Due to the lack of funding, early-phase translational research is often referred to as the "Valley of Death" in the drug development pipeline. In this chapter, we study an alternative financing model proposed by Fernandez et al. [23]—an RBO structure funded using both debt and equity—for early-stage orphan drug development. We extend their framework to account for dependence between phase transitions in projects, thus making it a more realistic representation of biopharma R&D. Using a multi-state, multi-period simulation approach, we characterize the performance of different megafund structures over a wide range of assumptions.

We find that our "vanilla" megafund demonstrates risk-reward characteristics attractive to both fixed-income investors and equity holders. The default risks of its debt tranches are comparable to the historical default rates of AAA-rated corporate bonds. In addition, the expected returns and the Sharpe ratio of the vanilla megafund are promising when compared to the CRSP index. Because R&D projects typically have small betas (i.e., weak correlation with market returns), the RBO structure can be an attractive option to investors seeking to diversify their portfolios away from conventional instruments. Consistent with previous studies, our results also show that the performance of the megafund can be further improved with the addition of a funding guarantee. Although the face value of the considered guarantee is large, the expected cost to the guarantor is, in fact, very small.

We simulate an equity-only structure as a baseline for comparison with the vanilla megafund, and find that the latter outperforms the former both in terms of ROE and research impact (quantified by the number of compounds successfully sold in phases 2 and 3). The disparity in performance can be attributed to the use of leverage in the vanilla megafund, which allows it to acquire a larger and more diversified portfolio. As shown in Table 6.3, equity returns generally increase with leverage in the capital structure. However, we note that adding leverage increases the volatility and default

risk of the megafund as well. Because of the trade-off between risk and return, greater leverage is not always better, and will therefore depend on the risk profile of the assets in the portfolio. The size of the debt tranches should be carefully selected to maximize the ROE while keeping the risk of default below thresholds acceptable to institutional investors.

In addition to the capital structure, we investigate the sensitivity of our results with respect to different project acquisition strategies, assuming a range of correlations and probabilities of success. We observe lower returns when the portfolio is constructed in stages over time instead of a single period at the start of simulation. This is explained by the projects acquired later having less time for development over the tenor of the megafund. In these cases, the use of more sophisticated securitization techniques like dynamic leverage can help to improve its performance [197].

In contrast with previous studies, we do not assume independence between phase transitions. The introduction of correlation leads to fatter tails in the distribution of returns, which imply higher probabilities of debt default and equity loss. However, we find that the senior tranche is protected by credit enhancement mechanisms from systematic risk even at high levels of correlation in the portfolio. In general, the vanilla and guarantee-backed megafunds outperform the all-equity structure over a wide range of correlations and probabilities of success.

We emphasize that our simulation is based on specific modeling assumptions regarding the cost, duration, valuation, and transition probability of clinical trials at each stage of development (outlined in Table 6.1). As seen in Tables 6.4 to 6.6, the expected performance of the megafund can change materially when different parameter values are used. The usefulness of our results depends heavily on the accuracy of the parameter estimates.

Unfortunately, given the nature of biopharma R&D, model calibration is especially challenging. For example, drug development projects are notoriously difficult to value since domain experts tend to have conflicting opinions on the therapeutic potential and the market value of investigational drugs. This is particularly common for first-in-class programs with novel treatment pathways. Furthermore, project outcomes are

often dependent on factors which cannot be easily quantified, e.g., the expertise and experience of the investigators and the managers in charge of the clinical trials.

Here, we use the empirical estimates proposed by Fagnan et al. [190] based on industry averages and expert panel evaluations for a rare disease portfolio. We also update the parameters for duration and phase transition based on a more recent study by Wong et al. [17] using two large pharmaceutical databases. Our assumptions may be considered conservative, since they do not account for possibilities that can make orphan drug development less costly or more lucrative, e.g., adaptive clinical trials that cost less and require shorter durations, or priority review vouchers (PRVs) that can be sold for additional revenue. (As an illustration, GW Pharmaceuticals received a PRV from the U.S. Food and Drug Administration for developing Epidiolex, a drug that treats rare childhood epilepsy. It sold the PRV to Biohaven Pharmaceutical for $105M in March 2019 [198].)

We should note that the investment mandate of the megafund outlined in this work is related to, but differs from, that of the "biopharmaceutical mega-fund" proposed by Ortiz et al. [199]. We consider the financing of a portfolio of risky early-stage pre-clinical assets, in contrast to their objective of securitizing a large pool of phase 1 assets. In addition, they investigate the potential benefits of incorporating assets backed by revenue-generating licensing and royalty agreements with well-capitalized entities.

Also, despite our focus on orphan drugs here, our framework can be easily generalized to arbitrary drug development portfolios once the simulation parameters are modified accordingly.

Chapter 7

Financing Treatments for Glioblastoma

Glioblastoma multiforme (GBM) is the most common type of brain cancer that is also the most aggressive and deadly. The prognosis for GBM is extremely poor, and one of the lowest among all cancers, with less than 6% of afflicted patients surviving more than five years after diagnosis. In fact, the survival rates for GBM have not shown any improvements over the past three decades. Treatment options are very limited. Apart from surgery and radiation, there are only four U.S. Food and Drug Administration approved drugs for brain tumors. Due to the significant scientific challenges, high costs of development, long investment horizons, and low probabilities of success, the development of curative treatments for GBM has largely remained stagnant. In this chapter, we investigate the use of a megafund [23, 189, 190, 200, 201, 202] as a financing vehicle to diversify and reduce the financial risks of drug discovery for GBM. We extend the simulation framework in previous studies to include adaptive clinical trial designs—specifically, the Glioblastoma Adaptive Global Innovative Learning Environment platform [203]—in addition to the traditional fixed-sample and fixed-duration protocols. We collaborate with the scientific team at the National Brain Tumor Society (NBTS) to identify a portfolio of 20 promising projects for investment, based on actual brain cancer therapies that are under development at the time of writing. Using modeling assumptions provided by the domain experts

from NBTS and from literature (e.g., probabilities of success, costs of development, durations, correlations, and profitability), we simulate the financial performance of the portfolio, and demonstrate that the megafund approach can provide promising returns to equity investors through diversification across different phases of development and multiple therapeutic mechanisms. Our results show that the megafund model has the potential to overcome current financial disincentives and accelerate innovation in the development of treatments for GBM.

7.1 Introduction

Glioblastoma (GBM) is the most common and the most lethal malignant primary brain tumor in the United States. It has an extremely poor prognosis, due to an unclear pathogenesis and a lack of curative treatments. A study in 2017 reported that GBM accounts for 47.1% of primary malignant brain tumor incidence in the U.S., and its five-year relative survival rate is only 5.5%, significantly worse than the survival rate for all malignant brain and central nervous system tumors combined, 34.9% [204]. Under the current standard of care, consisting of maximal surgical resection followed by chemoradiation [205], approximately 70% of GBM patients experience recurrence within one year of diagnosis, and the median survival time is merely 14.4 months [206].

Developing curative treatments for GBM is a social imperative. Nevertheless, it is financially risky, due to the long investment horizon and the low probability of success. In theory, the financial risks of early-stage GBM drug development could be mitigated via the "multiple shots on goal" strategy of a megafund vehicle [23]. Instead of placing its entire stake into a single asset, a megafund invests in a sizable portfolio of clinical assets diversified across development stages and therapeutic mechanisms. The risk-return performance of such a portfolio can be attractive to many private sector investors. Furthermore, the inherent parallelism of the approach greatly increases the chance of producing breakthrough life-saving therapies for presently incurable diseases.

The megafund vehicle was originally proposed to finance translational research in oncology, and it was subsequently adapted to specific disease areas such as orphan diseases [189], Alzheimer's disease [200], and ovarian cancer [202]. It is currently under consideration as a financing method by the National Brain Tumor Society (NBTS), the largest nonprofit organization in the U.S. dedicated to advancing innovative treatments of brain tumors with a vision to ultimately conquer and cure these deadly diseases once and for all.

In this chapter, we demonstrate the viability of applying the megafund vehicle to finance drug development programs for GBM and gliomas. Using estimates from the NBTS network of GBM experts and an extensive literature review, we perform Monte Carlo simulations to analyze the performance of such a megafund. We find that diversifying the portfolio across different stages of development and therapeutic mechanisms makes the risk-return profile acceptable to a large group of investors in the private sector. Furthermore, we demonstrate the synergy between the megafund approach and the novel platform clinical trial program Glioblastoma Adaptive Global Innovative Learning Environment (GBM AGILE) in simultaneously reducing the scientific and financial risks of developing early-stage innovative GBM therapies.

7.2 Parameters

In this analysis, we characterize the financial returns of a hypothetical portfolio of brain cancer therapeutics using Monte Carlo simulation. We adopt a framework similar to that used in prior studies to analyze the performance of an Alzheimer's disease megafund [200], a pediatric oncology megafund [201], and an ovarian cancer portfolio [202] (see Fig. 7-1). To reflect recent breakthroughs in brain cancer drug development, we extend the simulation model to include adaptive clinical trial designs in addition to the traditional fixed-sample protocol. In particular, we consider the GBM AGILE trial design [203]. As an inferentially seamless phase 2/3 platform trial [207], GBM AGILE has the potential to identify effective therapies for GBM more efficiently and rapidly than earlier methods. The cost and duration of such trials

are typically substantially lower than conventional clinical trials. Projects eligible for GBM AGILE significantly improve the risk-reward profile of the portfolio, which has special importance given the low historical success rates of treatment development for brain cancer [208].

This framework depends on a number of key modeling assumptions about the size and composition of the portfolio, the correlation between development outcomes, and the potential economic value of successful compounds. Each asset in the portfolio is assigned a probability of success, cost of development, and duration of clinical testing at each phase of development. We describe each aspect in detail in the following sections.

7.2.1 Portfolio

The performance of the megafund depends crucially on the composition of its underlying portfolio. To exploit the benefits of diversification and achieve an attractive risk-reward profile for the megafund, the portfolio should ideally cover a range of scientific pathways, mechanisms of action, and molecular targets, prudently allocating more capital towards projects that demonstrate strong scientific evidence, but also investing in programs based on more speculative hypotheses. In practice, project selection for the megafund would typically be performed by a team of medical experts and portfolio managers exercising scientific and business judgment and acumen acquired through years of domain-specific experience.

In this work, we identify scientifically promising pathways based on discussions with neuro-oncologists and leading industry experts from the NBTS network and the scientific team. This process yielded 20 projects for inclusion in our hypothetical NBTS portfolio (see Table 7.1). The projects are based on actual brain cancer therapies under development at the time of writing, spanning from assets in the late-stage discovery phase through the early- to mid-phases of clinical development. We also asked these experts to identify treatments that are potentially transformative, eligible for inclusion in GBM AGILE, or eligible for regulatory incentives, such as an Orphan Drug or Priority Review designation. This information is used to estimate

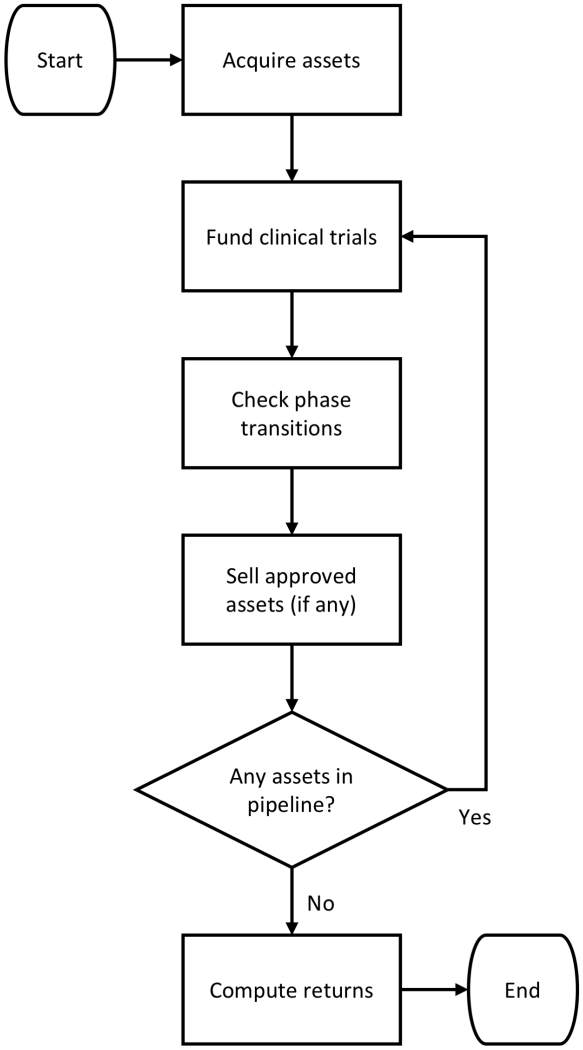


Figure 7-1: Simulation framework for the megafund. The fund acquires a portfolio of investigational assets at the start of the simulation. Pipeline drugs that successfully advance to the next phase of development are funded; those that fail are discontinued. Assets are liquidated at market value on approval. We compute the returns of the megafund at the end of the simulation.

the profitability of approved drugs.

Table 7.1: Hypothetical portfolio of brain cancer therapeutics. We assume that projects targeting pediatric patients are eligible for Priority Review Vouchers. *Abbreviations*: GA, eligibility for GBM AGILE; ODS, eligibility for Orphan Drug status; PP, target pediatric patients; TT, transformative treatment; IMM, immunotherapy; DDR, DNA damage repair; TM, tumor metabolism; PM, precision medicine; DE, devices; PRE, preclinical; P1, phase 1; P2, phase 2; HGGs, high-grade gliomas; uMGMT, unmethylated O6-methylguanine DNA methyltransferase; DNA-PK, DNA-dependent protein kinase; ATM, ataxia-telangiectasia mutated; ATR, ataxia telangiectasia and Rad3-related protein; FGFR, fibroblast growth factor receptor; LPCAT1, lysophosphatidylcholine acyltransferase 1; DRD2, dopamine receptor D2; BBB, blood-brain barrier; EGFR, epidermal growth factor receptor.

Therapeutic area	**Project**	**Patient population**	**Phase**	**GA**	**ODS**	**PP**	**TT**
IMM	T cell activation	Recurrent GBM	P2	Yes	Yes	No	Yes
	T cell activation	Recurrent GBM	P2	Yes	No	No	Yes
	T cell activation	Recurrent GBM	P2	Yes	Yes	No	Yes
	Personalized dendritic cell vaccine	Newly diagnosed GBM and HGGs	P1	Yes	Yes	Yes	Yes
	Retroviral replicating vectors	HGG	PRE	No	Yes	Yes	Yes
	Oncolytic virus	Recurrent GBM	PRE	Yes	Yes	No	Yes
	Autologous tumor cell vaccine	Newly diagnosed GBM	P2	Yes	Yes	No	Yes
DDR	DNA-PK inhibitor	Newly diagnosed uMGMT GBM	P2	Yes	Yes	No	Yes
	ATM inhibitor	Newly diagnosed uMGMT GBM	P2	Yes	Yes	No	Yes
	ATR inhibitor	Newly diagnosed GBM	P2	Yes	Yes	No	Yes
	FGFR inhibitor	Recurrent GBM	P2	Yes	Yes	No	Yes
	DNA repair inhibitors	Newly diagnosed uMGMT GBM	PRE	No	Yes	No	No
	ATM inhibitor	Pediatric gliomas	PRE	No	Yes	Yes	Yes
TM	LPCAT1 inhibitor	Newly diagnosed and recurrent GBM	PRE	No	Yes	No	No
PM	DRD2 receptor antagonist	Recurrent GBM with EGFR-low and DRD2-high tumor phenotype	P2	Yes	Yes	Yes	Yes
	BBB-penetrant signaling inhibitor	Newly diagnosed GBM	PRE	Yes	Yes	No	No
	CRISPR-Cas9 gene editing	Newly diagnosed and recurrent GBM	PRE	Yes	Yes	No	Yes
	BBB-penetrant transcription factor inhibitor	Newly diagnosed GBM	PRE	Yes	Yes	No	No
	BBB-penetrant transcription factor inhibitor	Brain metastases	PRE	Yes	Yes	No	No
DE	Fluorescence-guided surgery	Brain tumor	P2	No	Yes	No	No

7.2.2 Probability of Success, Cost of Development, and Duration

We first compile from the literature a set of estimates about the probability of success, the cost of development, and the testing duration of each phase of brain cancer drug development [189, 202, 208, 209, 210, 211] (see Appendix F.1). Next, we ask each expert from the NBTS network to estimate the same set of parameters based on their experience. To reduce the impact of outliers, we focus on the median of the estimates provided by the panel (see Appendix F.2). Finally, we take the average of both sets of estimates—those derived from the literature and the median of expert opinion—as the baseline values for our simulation (see Table 7.2).

Assuming standard clinical trials, we estimate that a brain cancer drug requires approximately 12 years of clinical development and about $110 million in development costs to move from preclinical stage to approval by the U.S. Food and Drug Administration (FDA). The baseline overall probability of success is estimated to be about 5.7%. This low figure reflects the challenges in developing brain tumor treatments, such as the lack of clinical trials for patients with brain metastases and the difficulties in delivering drugs across the blood-brain barrier, but it also implies that the unmet need and market potential in this patient population is very large.

We believe that a portfolio handpicked by the NBTS medical and scientific advisory council has the potential to outperform the industry average. Therefore, we adjust the overall probability of success estimate upward by a factor of 1.25x (a "skill and access factor" calibrated through discussions with the NBTS network of GBM experts) to 7.2% for our simulations. In the Section 7.3.2, we perform a sensitivity analysis of our results with respect to this factor.

7.2.3 GBM AGILE

GBM AGILE is a global, two-stage platform trial designed to facilitate the expedited approval of effective therapies for GBM, and reduce the cost of performing large-scale clinical studies [203]. It is operated by the Global Coalition for Adaptive Research

Table 7.2: Probability of success, costs of development, and duration at each phase of development for standard clinical trials. We believe that the NBTS portfolio has the potential to do better than the industry average. Therefore, we adjust the overall probability of success estimate upwards by a factor of 1.25x (a "skill and access factor"). This factor is distributed evenly among preclinical, phase 1, phase 2, and phase 3, i.e., an increase of approximately 1.06x for each phase, so that the overall probability of success from preclinical to approval is increased by 1.25x. *Abbreviations*: PoS, probability of success; PRE, preclinical; P1, phase 1; P2, phase 2; P3, phase 3; NDA, New Drug Application.

Parameter	**PRE to P1**	**P1 to P2**	**P2 to P3**	**P3 to NDA**	**NDA to Approval**	**PRE to Approval**
PoS (%)						
Baseline	64.5	70.7	35.3	35.8	100.0	5.7
NBTS portfolio	68.2	74.8	37.3	37.9	100.0	7.2
Skill and access factor	1.06x	1.06x	1.06x	1.06x	1.00x	1.25x
Duration (months)	12.0	33.1	39.3	50.0	10.8	145.1
Development cost ($ millions)	3.4[A]	8.3	18.6	81.4	0.0	111.7
Discount factor (%)	23.0	20.0	17.2	12.5	10.0	

[A] Includes an upfront cost of $2.3M.

(GCAR), a 501(c)(3) nonprofit organization. A platform trial evaluates the effects of multiple therapies, each as an experimental arm, against a common control arm. The platform is maintained under a master protocol, and therapies enter or exit the platform based on a decision algorithm [212]. In GBM AGILE, all drugs that enter the platform first undergo a screening stage ("stage 1"), which identifies promising therapies and enrichment biomarkers using overall survival as the primary endpoint. After a short burn-in period with fixed randomization to acquire initial response data, newly enrolled patients are assigned treatments via Bayesian adaptive randomization, with the probability of receiving each therapy proportional to the probability of that therapy improving overall survival. Promising therapies identified in the first stage then seamlessly transition to the second stage (the "confirmation stage" or "stage 2"), which uses fixed randomization on a smaller number of patients to confirm the therapeutic effects to support registration for FDA approval.

Under GBM AGILE, patient enrollment for arms that demonstrate promising results is prioritized and therefore, effective therapies proceed more rapidly through the trial, thus enabling faster registration. This can substantially reduce the cost and duration of developing GBM therapeutics. Therapies that do not enter the

confirmatory stage may still generate valuable clinical data for biopharma companies to improve drug and trial designs outside of GBM AGILE. Biopharma companies may also conduct follow-up trials—standard phase 2 or phase 3—for therapies that exhibit positive effects in stage 1, but do not meet the criteria to enter stage 2.

We model GBM AGILE as a two-stage process—stage 1 and stage 2—in place of the standard phase 2 and phase 3 trials (see Fig. 7-2). To simulate the uncertainty in the project selection process, we assume that each asset in the portfolio that is eligible for GBM AGILE has some probability of being included in the platform. Assets not selected for GBM AGILE proceed via the standard 505(b)(1) pathway for registration.

For assets in stage 1 of GBM AGILE, we assume that those which demonstrate promising treatment effects earlier will enter stage 2 with a smaller number of accrued patients ("early graduation")—further reducing the cost and duration of these trials. Other assets may either enter stage 2 after enrolling a larger number of patients ("regular graduation"), or exit the platform after stage 1 due to futility or tolerability issues. We also simulate the scenario where the megafund conducts follow-up standard phase 2 or phase 3 trials for assets that exit GBM AGILE after stage 1. Similar to the phase transitions of standard trials, we model inclusions in GBM AGILE and transitions from stage 1 and stage 2 as correlated Bernoulli random variables (see Section 7.2.4 and Appendix F.4).

We derive our cost and duration estimates assuming a steady state of one control arm and three experimental arms in GBM AGILE (see Appendix F.3 for baseline assumptions). We calibrate our estimates—patient accrual rate, cost per patient, and probability of inclusion and graduation of each stage—with the input from both NBTS and GCAR (see Table 7.3). No literature estimates were available at the time of writing since GBM AGILE is the first global, disease-specific platform trial for GBM. We note that the cost and duration of each GBM AGILE trial are much lower than standard phase 2 and phase 3 trials combined, approximately 75–85% lower in terms of cost, and 20–30% shorter in terms of duration.

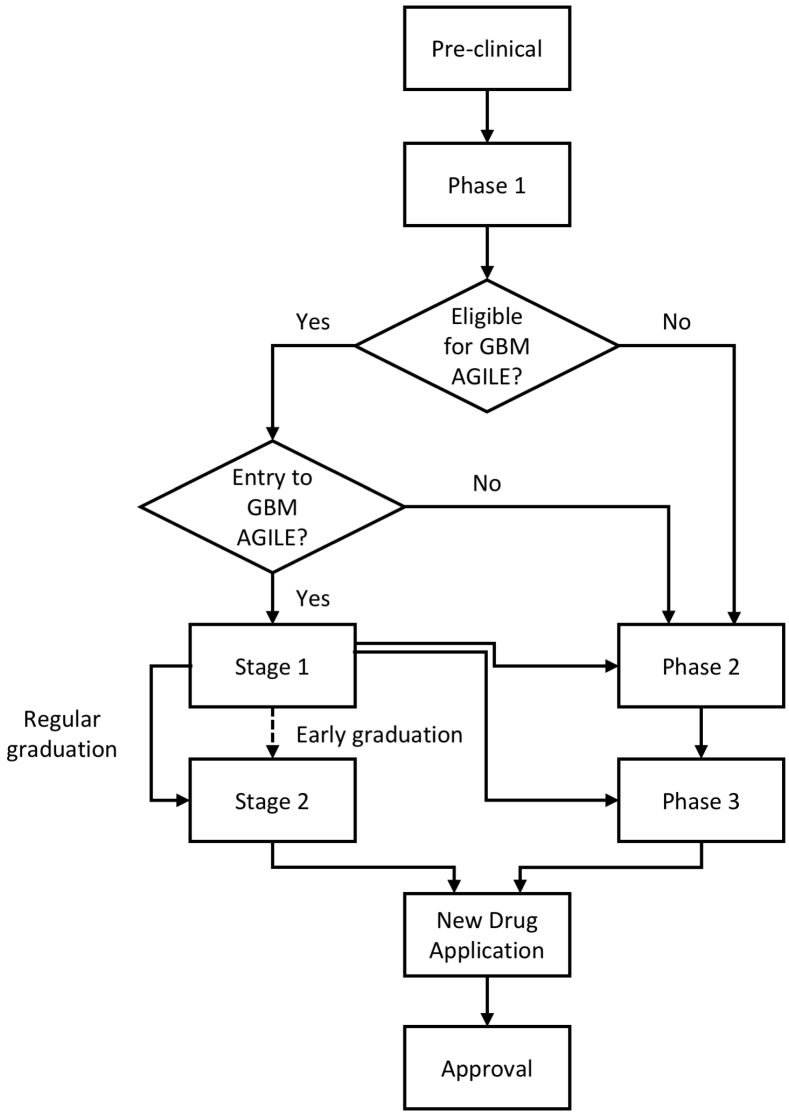


Figure 7-2: Possible development paths for assets in the NBTS portfolio.

Table 7.3: Probability of transition, costs of development, and duration at each stage of development for GBM AGILE. *Abbreviations*: PoT, probability of transition; P1, phase 1; S1, stage 1; S2, stage 2; P2, phase 2; P3, phase 3; NDA, New Drug Application.

Parameter	**P1 to S1**	**S1 to S2 (Regular)**	**S1 to S2 (Early)**	**S1 to P2**	**S1 to P3**	**S1 to Futility**	**S2 to NDA**
PoT (%)	33.0[A]	15.0	5.0	10.0	10.0	60.0	50.0
Duration (months)	7.0[B]	21.0	15.0	21.0	21.0	21.0	42.0
Development cost ($ millions)		15.2	10.7	15.2	15.2	15.2	7.5

[A] Probability of inclusion in GBM AGILE conditioned on a successful phase 1 trial (see P1 to P2 in Table 7.2). [B] Negotiation period for inclusion in GBM AGILE.

7.2.4 Correlation

The presence of pairwise correlation among the outcomes of therapeutic projects has major implications for the performance of the megafund. It introduces systematic risk to the portfolio that cannot be diversified away, and it has adverse effects on the risk profile of the fund in general. Depending on the similarities between the underlying treatment pathways and targets of projects in the portfolio, the outcomes of these projects (e.g., phase transitions, and entry to and graduation from GBM AGILE) are likely correlated with one another. That is, drugs with similar mechanisms of action are likely to have similar trial outcomes.

To quantify the level of correlation in our hypothetical portfolio, we asked the NBTS network of experts to estimate the pairwise correlation between every pair of projects in the portfolio. The correlations were first qualitatively assessed as low, low-medium, medium-high, and high by the team, and subsequently mapped to numerical values of 10%, 25%, 75%, and 90%, respectively. In the final step, we average the estimates by the experts (see Fig. 7-3) before projecting the resulting correlation matrix to its nearest positive-definite counterpart for use in our simulations [213]. See Appendix F.4 for the details of implementation.

7.2.5 Profitability of an Approved Compound

Brain cancer patients have very limited treatment options. The standard of care that has remained largely unchanged for over 20 years consists of surgery followed by radiation and temozolomide treatment. The other three FDA approved drugs for use in brain tumors, lomustine, carmustine, and bevacizumab, offer limited survival benefits. With so few historical data points, it is difficult to estimate the profitability of an approved brain cancer drug, as Chaudhuri et al. [125] did for ovarian cancer therapeutics. To complicate the process, the projects in our portfolio target a variety of patient populations, such as newly diagnosed patients, patients with recurrent disease, and adult versus pediatric patients. Therefore, it is quite likely they will have different valuations on approval. The use of a single market value as in Chaudhuri et

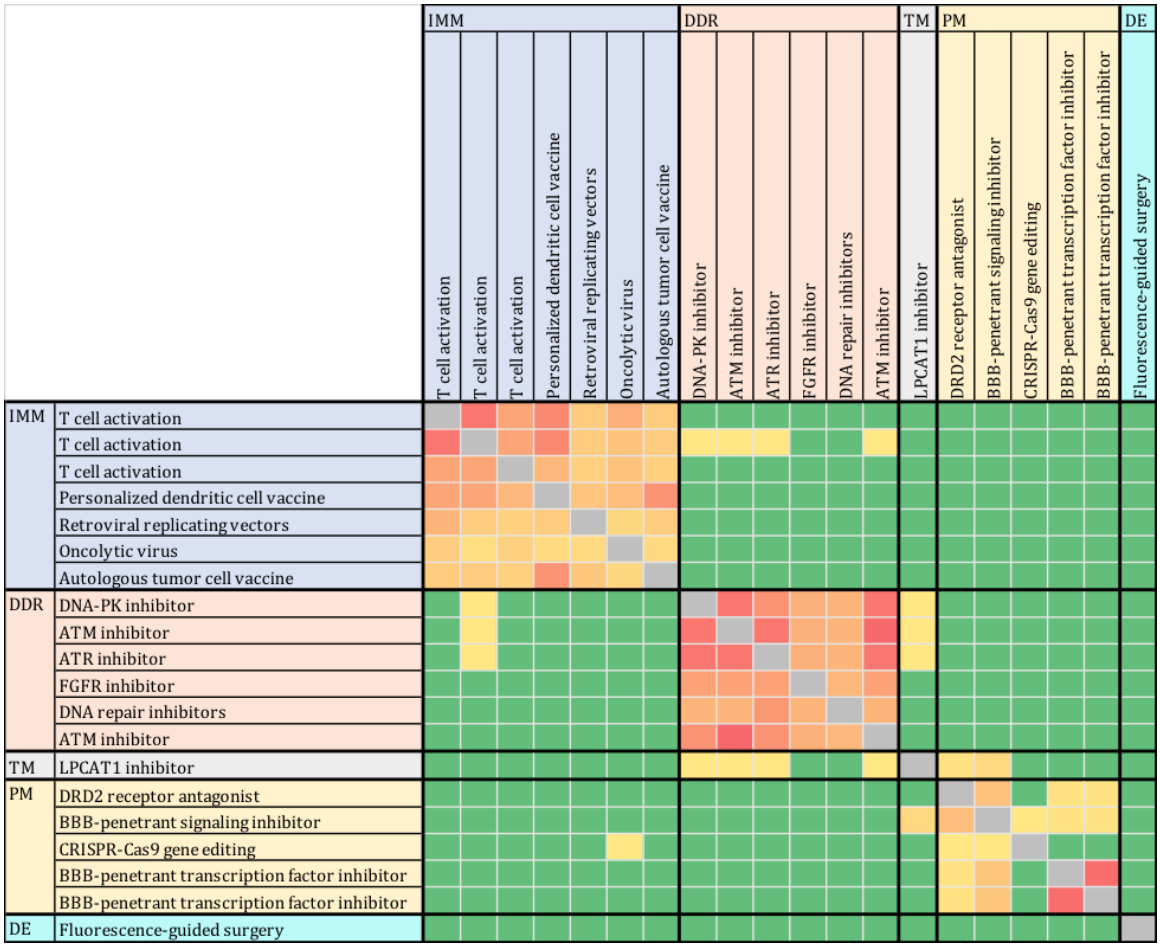


Figure 7-3: Correlation matrix of brain cancer projects (average of estimates from all the experts). Heat-map representation of pairwise correlations between all 380 ordered pairs of projects in our hypothetical portfolio. Red indicates high correlation, orange indicates medium-high correlation, yellow indicates low-medium correlation, and green indicates low correlation. *Abbreviations*: IMM, immunotherapy; DDR, DNA damage repair; TM, tumor metabolism; PM, precision medicine; DE, devices; DNA-PK, DNA-dependent protein kinase; ATM, ataxia-telangiectasia mutated; ATR, ataxia telangiectasia and Rad3-related protein; FGFR, fibroblast growth factor receptor; LPCAT1, lysophosphatidylcholine acyltransferase 1; DRD2, dopamine receptor D2; BBB, blood-brain barrier.

al. [125] may not be appropriate in this analysis.

Instead, we follow the approach used by Lo et al. [200] and Das et al. [201]. We estimate the economic value of a successful compound by the net present value (NPV) of its projected future cash flows upon FDA approval. The future cash flows are estimated using a set of assumptions about the incidence rate of the targeted patient population, the potential market penetration, the price charged per patient, the marketing exclusivity period, and the eligibility for pediatric extension and a Priority Review Voucher. In addition, we take into account the transformative potential of the treatment pathway; transformational treatments that substantially improve patient outcomes are priced at a premium—a "transformative factor" of 2.0x—relative to the standard of care. We calibrate our assumptions through discussions with the NBTS network of experts, and a review of the current standard of care and market research reports [214]. In our base case, the NPV of approved drugs in our portfolio ranges between $530M and $2,988M, with a median valuation of $1,272M. Fig. 7-4 illustrates the investment timeline of one drug in our portfolio. See Appendix F.5 for our baseline assumptions and valuation of all other drugs.

7.3 Results

7.3.1 Baseline

We summarize the performance of the NBTS portfolio in Table 7.4. The baseline portfolio achieves an average return of 15.1% per annum, which outperforms similar megafund portfolios for ovarian cancer [125] and Alzheimer's disease [200], suggesting that such an investment opportunity may be attractive a wide group of private sector investors. Its NPV is approximately $86M, indicating that the megafund is likely to be profitable.

On the other hand, the baseline portfolio demonstrates high volatility and large probabilities of loss and wipeout, a limitation imposed by the scientific challenges of GBM therapeutic innovation. Nonetheless, our simulation shows that, on average,

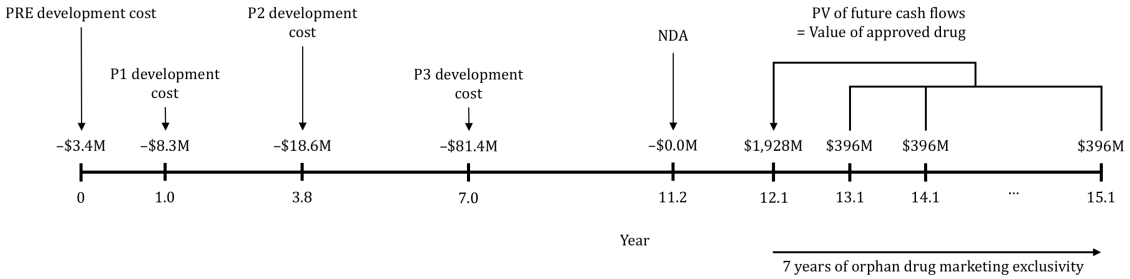


Figure 7-4: Investment timeline of a brain cancer drug targeted at recurrent GBM patients. We assume an incidence rate of 30,000 patients per year, a conservative market penetration of 10%, and a price of $132,000 per patient. For simplicity, we assume that our price is the amortized cost of the entire course of treatments needed for each patient. We use the annual per-patient expenditure of Temodar—computed based on the average wholesale price—as reference for the price of a newly approved GBM drug. We believe that a new therapy with greater efficacy over the standard of care and marketing exclusivity is likely to be priced closer to a brand-name drug than a cheaper generic drug like temozolomide. The annual cost of Temodar is estimated to be $66,000 per patient [215], adjusted to 2019 dollars using the Biomedical Research and Development Price Index. We note that the drug in this example is priced at a premium of 2.0x, i.e., $132,000, because it has been identified to be a transformative treatment by the experts. Assuming a cost of capital of 10%, the drug has a net present value at $1,928M on approval. *Abbreviations*: PRE, preclinical; P1, phase 1; P2, phase2; P3, phase 3; NDA, New Drug Application; PV, present value.

more than two therapies financed by the megafund will receive FDA approval, though the uncertainty in the number of approved therapies is also significant. There is a 79.1% probability that at least one therapy in the portfolio will receive FDA approval, and the average duration from the initial acquisition of the assets until the first FDA approval is 8.3 years. The advent of new curative therapies for GBM should generate tremendous societal benefits on a global scale, which cannot be captured by these financial metrics.

Table 7.4: Performance of the NBTS portfolio, simulated with 100,000 Monte Carlo runs. The baseline portfolio consists of 12 projects from a mix of phases, 15 of which are eligible for inclusion in GBM AGILE, and assumes a skill and access factor of 1.25x, a transformative factor of 2.0x, a maximum market penetration of 20% for approved drugs, and a correlation matrix derived from estimates by the NBTS network of experts. E[·] denotes expectation and SD[·] denotes standard deviation. *Abbreviations*: R, average return per annum; N, number of approved drugs; T, time until first FDA approval; PoL, probability of loss; PoW, probability of wipeout; ρ, pairwise correlation; SAF, skill and access factor; TF, transformative factor; MP, rate of market penetration.

	E[R] (%)	**SD[R] (%)**	**E[NPV] ($ millions)**	**SD[NPV] ($ millions)**	**E[N]**	**SD[N]**	**E[T] (years)**	**SD[T] (years)**	**PoL (%)**	**PoW (%)**
Baseline	15.1	24.3	86	780	2.2	2.0	8.3	1.7	25.6	20.9
Preclinical only	11.4	26.3	-19	398	1.5	1.6	11.5	0.9	37.2	33.8
Equi-correlation ($\rho = 0\%$)	17.4	18.6	80	574	2.2	1.4	8.2	1.6	14.3	9.5
Equi-correlation ($\rho = 10\%$)	16.1	21.5	81	673	2.2	1.7	8.2	1.6	19.9	15.1
Equi-correlation ($\rho = 40\%$)	12.4	29.2	91	962	2.2	2.5	8.2	1.6	34.7	30.2
Equi-correlation ($\rho = 80\%$)	4.4	42.7	86	1,419	2.2	3.8	8.1	1.5	56.6	53.6
SAF = 1.0x	12.8	25.0	16	737	1.9	1.9	8.3	1.7	29.3	24.9
TF = 1.0x	8.7	19.3	-60	434	2.2	2.0	8.3	1.7	31.5	20.9
MP = 10%	6.4	17.5	-95	376	2.2	2.0	8.3	1.7	32.8	21.2
MP = 30%	18.2	27.0	170	1,185	2.2	2.0	8.3	1.7	24.3	20.8
Standard clinical trials only	11.6	22.3	-38	711	2.2	2.0	8.9	1.3	28.5	22.9

7.3.2 Sensitivity Analysis

To analyze the robustness of our results against each model assumption, we perform sensitivity analyses on the portfolio composition, the correlation structure between projects, the added value of biomedical expertise, the price premium of transformative therapies, as well as the potential market penetration of an approved drug. We summarize the results in Table 7.4.

Portfolio Composition

The performance of the portfolio hinges on its diversification. In addition to the baseline portfolio which consists of projects from a mix of phases (see Table 7.1), we simulate a portfolio with the same set of projects but all acquired at the preclinical stage. This allows us to gauge the effect of diversification across different stages of development.

We find that a preclinical-only portfolio requires an average investment of only \$673M, much lower than the \$1,037M of the mixed-stage baseline portfolio, since market valuations are based on lower probabilities of success and longer investment horizons. However, the lack of diversification across development stages substantially increases the risk that no therapy in the portfolio will receive FDA approval, leading to a 3.7 percentage point fall in expected annualized return and an 11.6 percentage point increase in the probability of loss versus the base case. The expected NPV also becomes negative, indicating that the investment will result in a net loss. To ensure an attractive risk-return profile, it is critical to structure the portfolio with assets spanning early- and mid-phases of clinical development.

Correlation

The volatility of the portfolio is largely determined by the correlation structure of the underlying drug development programs. A portfolio consisting of multiple drugs that are highly correlated with one another, e.g., based on similar therapeutic mechanisms, will have high volatility. To test the sensitivity of our results to this parameter, we

simulate portfolios where the pairwise correlation between any two distinct assets is identical, i.e., equi-correlation. We sweep values between 0% and 80% and find that the expected annualized return decreases with higher correlation, while all other risk measures—probability of loss and wipeout, and volatility of returns—increase.

The correlation structure of the baseline NBTS portfolio is based on the qualitative assessment of program similarity by domain experts. Although certain groups of drugs in the portfolio are highly correlated due to similar therapeutic mechanisms, diversification across therapeutic mechanisms lowers the overall correlation to a level equivalent to that of an equi-correlated portfolio with pairwise correlations between 10% and 40%.

Skill and Access Factor

In our baseline scenario, we assume that the portfolio managers are skilled at identifying promising drug candidates for investment, and model this advantage by adjusting the overall probability of success estimate upwards by a skill and access factor of 1.25x. Here, we reduce the factor to 1.0x, implying no incremental improvement in the probability of success beyond the industry average. This reduces the expected annualized return by 2.3 percentage points, while increasing the probabilities of loss and wipeout by 3.7 and 4.0 percentage points, respectively. The expected NPV remains positive but decreases to less than a fifth of the baseline value. The sensitivity of the megafund performance to the skill and access factor reveals the importance of biomedical expertise in project selection and portfolio management.

Transformative Factor

In our base case, we assume that transformative therapies, once approved, can be priced at a premium of 2.0x relative to the current standard of care, thus generating much higher revenue than an average drug for GBM. As drug prices are facing increasing scrutiny from regulators, payers, and patients, such a high premium may be inappropriate and unrealistic. Therefore, we consider the case where all approved therapies are priced at the current standard of care, i.e., a pricing transformative

factor of 1.0x. We find that this leads to a 6.4-percentage-point decrease in expected annualized return, and a 5.9-percentage-point increase in the probability of loss. Furthermore, the expected NPV becomes negative, indicating that the ability and flexibility to price transformative therapies at a premium has substantial implications on the financial viability of a GBM venture fund.

Market Penetration

A key factor determining profitability and returns is the potential market penetration of an approved drug, i.e., the proportion of the target patient population who will receive the therapy once it enters the market. Our baseline model assumes the maximum market penetration of any approved drug to be 20%. This estimate is likely conservative, since no curative treatment of GBM is currently available. However, we can reasonably expect a transformative therapy to become the new standard of care for GBM once approved. Such a drug will likely acquire a market share well above 20%. Increasing the maximum market penetration to 30% increases the expected annualized return by 3.1 percentage points, and doubles the expected NPV. In contrast, decreasing the maximum market penetration to 10% reduces the expected annualized return by more than half, and turns the expected NPV negative. The impact of the market penetration on returns illustrates the significant profit potential of transformative GBM therapies.

GBM AGILE

The megafund vehicle and GBM AGILE share the same concept of "multiple shots on goal." They have complementary goals: the former facilitates the financing of multiple drug development programs in parallel, while the latter expedites the clinical investigation of multiple experimental treatments simultaneously. The baseline portfolio includes 15 assets that are eligible for inclusion in GBM AGILE, out of a total of 20 projects. To demonstrate the synergy between these two novel models, we simulate a portfolio with the same set of projects but without GBM AGILE, i.e., a portfolio with non-adaptive, conventional clinical trials only. In the absence of GBM

AGILE, we find that the expected annualized return falls by 3.5 percentage points. Furthermore, the expected NPV becomes negative, indicating that the venture fund will likely make a net loss. This illustrates the value added by GBM AGILE, in terms of boosting returns and NPV, and reducing risks.

7.4 Discussion

The development of new transformative therapeutics for GBM has been largely unsuccessful for decades. This is due not only to the inherent scientific challenges of development, but also the significant financial risks of investing in early-stage clinical programs. The performance of a GBM megafund may be attractive to a wide group of investors from both the public and private sectors, provided the underlying portfolio is suitably diversified and uses the GBM AGILE platform to increase its overall probability of approval.

Sensitivity analysis reveals that domain expertise plays a crucial role in identifying promising and potentially transformative therapies for investment in the GBM megafund portfolio. It also demonstrates the importance of diversification across different stages of development and mechanisms of action as an effective way to increase the megafund's overall probability of success and reduce the volatility of its returns.

In addition, the use of the novel GBM AGILE platform generates significant synergy with the megafund. Inclusion of portfolio assets in the platform boosts annualized returns and NPV, reduces risks, and expedites the delivery of transformative GBM therapies to patients, making the venture fund attractive to both private sector and impact investors. The GBM AGILE platform also provides a financially efficient means to collect valuable clinical data for a therapeutic asset to guide its subsequent development in clinical trials, even if the therapy does not meet the criteria to enter stage 2 of the platform.

In our simulations, we assume that enough capital can be raised to finance the entire portfolio through all stages of development. In practice, it may be difficult for nonprofit organizations such as NBTS to raise nearly $1.5B at the outset. To address

this issue, the fund may consider a mixture of equity and debt in its capital structure and adjust the leverage dynamically as the clinical trials progress into later stages [197]. Under a tight budget constraint, it may also be necessary to acquire drug development programs dynamically, liquidating some projects during intermediary development in order to fund more promising candidates. Our simulation results may be regarded as an upper bound on the performance of a GBM megafund in practice.

7.5 Conclusion

Developing curative treatments for GBM is an urgent social imperative. However, the high development costs, long investment horizons, and significant risks of failure in the clinical trial process have prevented private sector investors from investing in early-stage GBM drug development programs to treat this deadly disease. We demonstrate the potential viability of the megafund vehicle to finance a portfolio of 20 GBM drug development programs. Through the appropriate diversification of the portfolio across different stages of development and therapeutic mechanisms, while simultaneously leveraging the novel GBM AGILE platform to lower costs and accelerate clinical testing, the risk-reward profile of such a megafund can be attractive to equity investors.

Part IV

Conclusion

Chapter 8

Summary of Findings

Despite the many breakthroughs in biomedical research and the increasing demand for new drugs to treat unmet medical needs, the productivity of the pharmaceutical industry has not grown at the same pace. It appears that the traditional sources of financing in biopharma research and development (R&D) are no longer compatible with the new realities of biomedical innovation, a process which has become more challenging, complex, expensive, time-consuming, and risky in the past two decades. There is a need for better analytics in different areas of biomedical research to allow stakeholders to make more informed decisions. In addition, new business models are required to address the dearth of funding for translational medicine in the valley of death due to the mismatch between the risk characteristics of biomedical projects and the risk preferences of biopharma investors. We explore both topics in this thesis, with Chapters 2 to 4 in Part II focusing on the former and Chapters 5 to 7 in Part III on the latter.

In Chapter 2, we develop predictive analytics for precision medicine in advanced non-small cell lung cancer (NSCLC) that reflects the current standard of care. Our work is one of the largest studies of NSCLC to consider biomarker mutation and inhibitor therapy as candidate predictive variables. We propose a stochastic tumor growth model to predict tumor response, and consider a range of machine learning algorithms and survival models for predicting clinical endpoints. We estimate and validate our models using data from pivotal randomized clinical trials submitted to

the U.S. Food and Drug Administration, and demonstrate that our models achieve promising out-of-sample predictive performances. In addition, we identify baseline variables that are strongly associated with response and survival, and find that our results are consistent with related studies in the literature. Our findings also point to the need for data on composite multi-omic signatures in order to develop more powerful predictive models.

In Chapter 3, we develop better analytics for quantifying the risks in drug development projects. Using drug and clinical trial data from two large pharmaceutical pipeline databases, we train machine learning models to predict the probability of approval of drug candidates in phase 2 and phase 3, respectively. Unlike related studies that use complete case analysis—i.e., dropping all data points with missing information—we apply statistical imputation methods to deal with missingness. To the best of our knowledge, our analysis is the largest of its kind. The dataset used includes more than 6,000 unique drugs and over 19,000 unique clinical trials. In contrast, most published research are based on datasets with less than a hundred drugs. In addition, our models provide conditional estimates of success based on a wide range of drug and clinical trial features. Such estimates are more accurate than the unconditional estimates that biopharma companies typically use to manage their portfolios. Our models may be used to make more informed data-driven decisions in risk assessment and portfolio management. By reducing the uncertainty surrounding drug development and providing greater risk transparency, our work can help improve financial efficiency and facilitate capital allocation in biomedical research.

In Chapter 4, we propose a systematic framework for quantitatively assessing the costs and benefits of different vaccine efficacy clinical trial designs for COVID-19 (coronavirus disease 2019) vaccine development, including fixed-duration clinical trials, a novel adaptive clinical trial, and a human challenge trial (HCT). We use epidemiological models calibrated to the current pandemic to simulate the time course of each trial design for different vaccine efficacies, epidemiological scenarios, vaccination schedules, and approval requirements, and identify situations where HCTs can provide greater social value versus non-challenge trials. To the best of our knowl-

edge, our work is the first study that attempts to quantify the potential benefits of a COVID-19 HCT. Our results contribute to the moral and ethical debate about HCTs by allowing stakeholders, such as vaccine developers, policymakers, and HCT volunteers to understand the implications of their actions (or inaction).

In Chapter 5, we develop a systematic framework for tracking the outcomes of university technology licensing in the life sciences using the Massachusetts Institute of Technology as a case study. We construct several measures of impact including patents cited in the Orange Book, capital raised, outcomes from mergers and acquisitions, patents granted to university intellectual property licensees, drug candidates discovered, and U.S. drug approvals. As academic institutions play an increasingly important role in the biotechnology industry through technology transfer and the creation of startups, our methodology provides a useful framework for other institutions to track the outcomes of their intellectual property in the therapeutics domain. Our results also raise the possibility of a novel business model—an Academic Translational Medicine fund that raises capital from limited partners to invest in therapeutics companies that license intellectual property from a consortium of universities for further development and commercialization.

In Chapter 6, we investigate the use of a recently proposed megafund structure for financing early-stage biomedical research. We extend the existing model to account for technical correlation between assets in the underlying portfolio using a single-factor model with a Gaussian copula, thus allowing us to evaluate the tail risks of the megafund more accurately. We show that financial engineering techniques such as portfolio theory and securitization can be used to structure the megafund into derivatives (e.g., bonds and equity) with risk-reward characteristics that are attractive to different classes of investors. This allows the fund to tap into a substantially larger pool of capital (e.g., the bond market) than the traditional sources of biopharma R&D funding (e.g., venture capital, philanthropic donations, and public and private equity). By improving the efficiency of the financing process and lowering the cost of capital, the megafund approach can help alleviate the valley of death in translational R&D.

In Chapter 7, we collaborate with physicians and researchers at the National Brain Tumor Society to identify and model a portfolio of promising glioblastoma (GBM) therapies for investment. We further extend the megafund model to include the Glioblastoma Adaptive Global Innovative Learning Environment (GBM AGILE), a novel adaptive clinical trial platform designed specifically to accelerate clinical testing for GBM. Despite the low historical success rates of treatment development for brain cancer, our simulations show that a GBM venture fund can generate promising returns to equity investors through appropriate diversification. Our results also highlight the synergy between the GBM AGILE platform and the megafund approach: the former expedites the clinical testing of multiple experimental treatments simultaneously, while the latter facilitates the financing of multiple drug development programs in parallel. By leveraging on innovations in clinical trial design and financing mechanism, we can overcome current financial disincentives and accelerate the development of treatments for GBM, a deadly disease with very limited treatment options and huge unmet need.

Part V

Appendices

Appendix A

Supplement to Chapter 2

A.1 Scaling RECIST Measurements

Under RECIST, investigators select target lesions to follow throughout a patient's treatment course in the clinical trial. They use the sum of longest diameter in these target lesions (SLD) as a measure of tumor burden, and ultimately for response determination. When first published in 2000, RECIST (version 1.0) accommodated up to ten measurable target lesions. However, an update to the criteria in January, 2009 (version 1.1) reduced the maximum number from ten to five. As a result, there are two groups of SLD measurements in the dataset. Five of the studies in the dataset were initiated before RECIST version 1.1 was published and have SLD measurements based on more than five lesions. This group of measurements are, in general, greater than those from the remaining 12 trials that use the revised RECIST.

In order to reconcile SLD measurements collected under the older criteria with the current version, we scale earlier measurements to reflect the new five lesion limit by right-censoring measurements with more than five lesions to values equal to five times the size of the corresponding average target lesion. For example, a baseline SLD of 140 mm for seven identified target lesions has an average lesion size of 20 mm. Since it has more than five lesions, we scale the measurement to $5 \times 20 = 100$ mm. We keep measurements with less than five lesions as they are since they do not violate the new target lesion maximum. We also do not modify response outcomes

determined under RECIST version 1.0 because they are largely dependent on relative changes. We believe these adjustments achieve a good balance between discarding a considerable part of the dataset and distorting the data distribution. It is impossible to account for all changes introduced in RECIST version 1.1 perfectly due to the lack of precise patient-level tumor information and domain expertise.

A.2 Case Studies of Longitudinal Data

Clinical trials typically record patient-level SLD measurements collected at three types of visit, namely the baseline visit, the treatment visits, and the follow-up visits. The baseline visit of the clinical trial is self-explanatory, typically part of the screening visit for the trial. Treatment visits are conducted throughout the clinical trial, as patients receive the allocated study medication. These visits are scheduled at regular intervals, with the exact interval depending on the study protocol and patient availability. It is not uncommon for patients to miss visits, or for investigators to conduct additional unscheduled visits to confirm response or progression. At any point in the trial, patients can be withdrawn from treatment if any of the following occurs: documentation of PD, protocol violation, a serious adverse event, or their withdrawal of consent. While investigators typically follow these discontinuation rules closely, there is some discretion involved. For example, they may choose to treat beyond PD if they determine the subject is stable and deriving clinical benefits from the drug.

Follow-up visits are performed after the discontinuation or completion of the trial treatment. The number of follow-up visits depends on multiple factors. Some protocols schedule only one follow-up, while others may require investigators to track patients who discontinue for reasons other than PD until they experience PD or start a new anti-cancer therapy[1]. In some cases, follow-up visits are not possible, either because the subject withdrew consent, or became unreachable due to other reasons (e.g., death, relocation, etc). There are also other complicated scenarios, such as crossovers to the experimental therapy upon disease progression. We exclude such data to ensure that the effects observed are solely due to the therapies we are interested in, and not caused by other interventions.

We examine three case studies in detail in Fig. A-1 to illustrate some of the subtleties present in the longitudinal SLD dataset. Patient 1 represents a typical responder who demonstrates OR at some point in the trial. In this example, the re-

[1]The time course of response to treatment is typically delayed relative to drug administration [216, 217]. Therefore, follow-up visits prior to the start of another anti-cancer therapy are arguably as important as treatment visits, because drug effects may manifest themselves during this period of time.

sponse lasted for 18 consecutive weeks and is considered to be confirmed, barring any negative changes in non-target lesions or appearances of new lesions. The time series ends when PD in SLD occurs and the patient is discontinued. Patient 3 represents a typical progressor—that is, a patient who exhibits disease progression—whose SLD increases monotonically until PD is documented. Thereafter, the patient is taken off therapy without any follow-up visits. The lack of post-discontinuation measurements suggests that the patient either started a new therapy almost immediately, has been lost to follow-up visits, or simply withdrew consent. In similar cases of disease progression, however, follow-up visits may be present. Patient 2 presents a more interesting example. The patient demonstrates PD in SLD at week 12 of the trial, but unlike Patient 3, is kept on treatment, perhaps because the patient has been determined to derive clinical benefits despite growth. In contrast with Patients 1 and 3, the measurements for Patient 2 ended on SD, and not PD, at week 36. There are several possible explanations: Patient 2 could have been discontinued either because of progression in non-target lesions, the appearance of new lesions, serious adverse events preventing further administration of treatment, withdrawal of consent, or severe non-compliance with the study protocol. On the other hand, the patient could still be under therapy, and the lack of measurements is due to a data cutoff.

Apart from straightforward cases like Patients 1 and 3, it is often difficult to glean the exact reasons for discontinuations in tumor assessment from the dataset. Nevertheless, it is clear that there is a discontinuation process at work that affects our observation of SLD measurements. For example, we are less likely to observe measurements after PD in SLD has occurred. This phenomenon is in inherent to the data collection process because of the patient safety protections designed into the clinical trials.

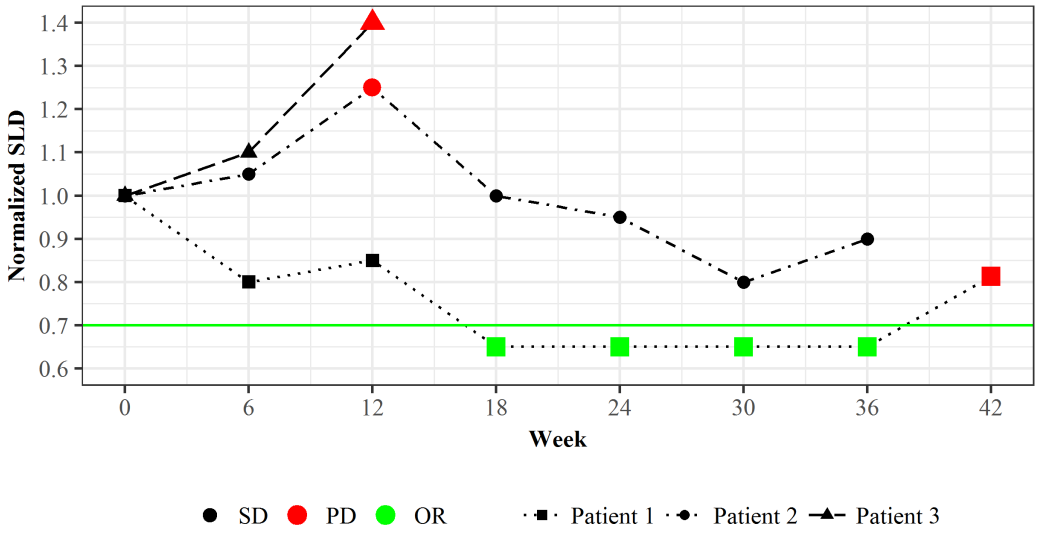


Figure A-1: Case studies of longitudinal SLD data for three example patients. We normalize the SLDs with respect to the corresponding BSLDs, and assume that visits are scheduled at intervals of 6 weeks. Timepoint SLDs that qualify as response or disease progression according to RECIST are highlighted in green and red, respectively.

A.3 Features for Predictive Models

Table A.1: List of predictive factors for tumor response. We have 45 categorical variables and 5 continuous variables. Categorical variables are converted to binary variables; continuous variables are centered and scaled by their standard deviations.

Type	**Variable**	**Values**
Treatment group[1]	Biomarker status and therapy received	Six categories[3]
Biomarker positivity[2]	PDL1, EGFR or ALK	Yes, no
Demographics	Age	Years
	Weight	kg
	Sex	Male, female
	Race group	Asian, white, others
	Region	APAC, NAM, WEUR, others
Medical history	Time since diagnosis	Days
	Performance status	0, 1, 2 or higher
	Smoking status	Ever, never
	Stage at screening	IIIB or lower, IV
	Prior chemotherapy	Yes, no
	Histology	Adeno, SCC, others
	Number of baseline target lesions	1, 2, 3, 4, 5 or more
	BSLD	mm
Metastasis	Brain	Yes, no
	Bone	Yes, no
	Liver	Yes, no
	Others	Yes, no
	Number of metastasis sites	Count
Comorbidities	Neoplasms benign, malignant and unspecified	Yes, no
	Immune system disorders	Yes, no
	Endocrine disorders	Yes, no
	Musculoskeletal and connective tissue disorders	Yes, no
	General disorders and administration site conditions	Yes, no
	Metabolism and nutrition disorders	Yes, no
	Vascular disorders	Yes, no
	Nervous system disorders	Yes, no
	Injury, poisoning and procedural complications	Yes, no
	Hepatobiliary disorders	Yes, no
	Eye disorders	Yes, no
	Respiratory, thoracic and mediastinal disorders	Yes, no
	Cardiac disorders	Yes, no
	Ear and labyrinth disorders	Yes, no
	Skin and subcutaneous tissue disorders	Yes, no
	Blood and lymphatic system disorders	Yes, no
	Reproductive system and breast disorders	Yes, no
	Congenital, familial and genetic disorders	Yes, no
	Infections and infestations	Yes, no
	Gastrointestinal disorders	Yes, no
	Social circumstances	Yes, no
	Psychiatric disorders	Yes, no
	Renal and urinary disorders	Yes, no
Laboratory measurements	Alkaline phosphate	High, normal, low
	Alanine aminotransferase	High, normal, low
	Aspartate aminotransferase	High, normal, low
	Bilirubin	High, normal, low
	Creatine	High, normal, low
	Hemoglobin	High, normal, low
	Platelets count	High, normal, low
	White blood cells count	High, normal, low

[1] For tumor response and PFS models. [2] For OS models.

[3] Any biomarker status and under chemotherapy, PDL1-positive and under PDL1 ICI, EGFR-positive and under EGFR TKI, ALK-positive and under ALK TKI, negative biomarker status but under inhibitor therapy (either PDL1-negative but under PDL1 ICI or EGFR-negative but under EGFR TKI) and not tested for any biomarkers but under inhibitor therapy (either PDL1 ICI or EGFR TKI).

Appendix B

Supplement to Chapter 3

B.1 Data Preprocessing

We construct our datasets from two Informa® databases: Pharmaprojects and Trialtrove, two separate relational databases organized by largely different ontologies. We extract drug-specific features and drug-indication development status from Pharmaprojects, and clinical trial features from Trialtrove. We merge the databases through keys provided separately by Informa®.

Pharmaprojects was created earlier than Trialtrove, and thus the disease coverage for clinical trials is not as extensive. We start the merging process by first identifying all drug-indication pairs in Pharmaprojects. Subsequently, we drop pairs that do not have any trials recorded in Trialtrove. As highlighted in Section 3.2.2, profiles in Pharmaprojects and Trialtrove are fraught with missingness. Therefore, we impose several filters when constructing the datasets to ensure that all instances collected are usable for analysis.

Table B.1 summarizes the steps in the filter. We note that the drug, indication, and trial relationships in the constructed datasets are surjective and non-injective: different drugs may target the same indication, and some trials may involve multiple drug-indication pairs. This is logical because it is common that drugs treat multiple diseases, multiple drugs treat a specific disease, or trials involve two or more related primary investigational drugs. To provide some intuition for the size of these

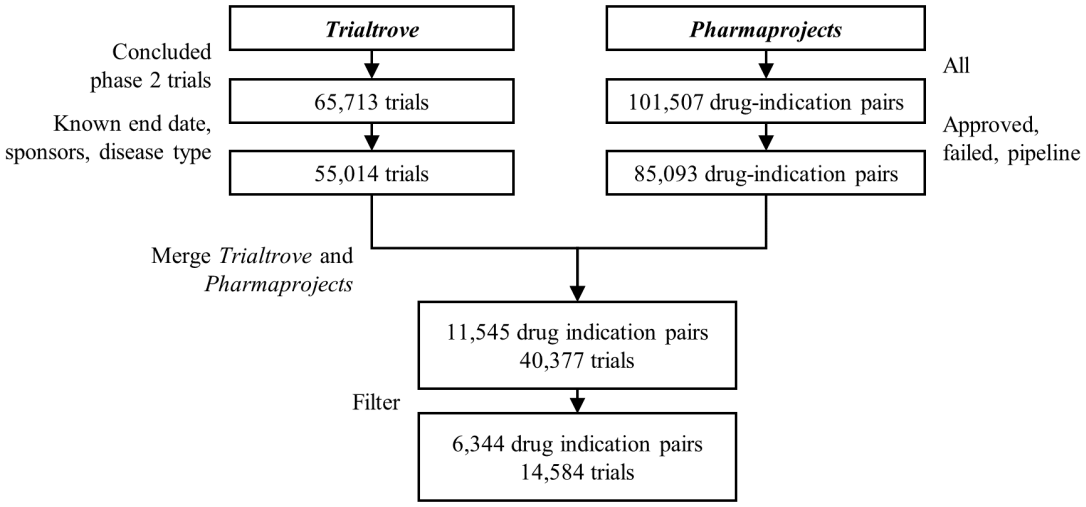


Figure B-1: Data cleaning for P2APP.

databases, we summarize, in Fig. B-1 and Fig. B-2 (for P2APP and P3APP, respectively), how the number of drug-indication pairs and clinical trials change as we perform the filters.

We extract drug compound attributes and clinical trial characteristics from Pharmaprojects and Trialtrove, respectively (see Table 3.1 and Table B.2). In addition to features readily available in the databases, we create an augmented set of variables capturing sponsor track record and investigator experience. We quantify the track record of sponsors of a specific trial by their success in developing other drugs, using the number of prior approved and failed drug-indication developments; and in past trials for phases 1, 2, and 3 separately, using the total number of trials sponsored, the number of trials sponsored with positive and negative results, and the number of trials sponsored to completion and termination. We use the end date of the last trial of the drug-indication pair under consideration as the cutoff for considering prior experience. This is because the last end date will be the time of prediction. We abstract investigator experience in the same manner. Lastly, we construct a binary drug-indication pair feature, whether the drug has been approved for another indication before. Similarly, we use the end date of the last trial as cutoff for considering prior approval. In total, our datasets have 31 drug-related features and 113 trial-related features.

Table B.1: Filters for creating datasets.

	Rationale
Drug-indication Pairs in Pharmaprojects	
Trials observed in Trialtrove (phase 2 for P2APP; phase 3 for P3APP)	We exclude pairs for which we do not observe any trials in Trialtrove.
Known approval date (if approved). Approval dates are not available directly in Pharmaprojects. They are embedded within text blocks. We had to mine these text blocks (combination of heuristics and manual extraction) to extract the dates.	We define the approval date as the earliest date a drug-indication pair was approved in any market. We need these dates to create an augmented set of variables capturing sponsors and investigators experience, and also to perform time-series analysis.
Known failure date (if failed)	Failure dates are not directly available in Pharmaprojects. We define failure date as one year after the end-date of the last phase 2 or phase 3 trial (if any), whichever is latest.
Clinical Trials in Trialtrove	
Phase 2 for P2APP; phase 3 for P3APP	We are interested in predicting approvals using trial features.
Known end date	We need these dates to perform time series analysis. For approved drug-indication pairs in P2APP and P3APP, we compare the trial end date with the corresponding approval date to filter out post-approval trials. These trials may be for supplemental New Drug Applications (e.g., modified dosage) that are irrelevant to our analysis.
Known sponsors and disease types	Trials not tagged with sponsor/disease types are typically out of Trialtrove commercial coverage and are not maintained.

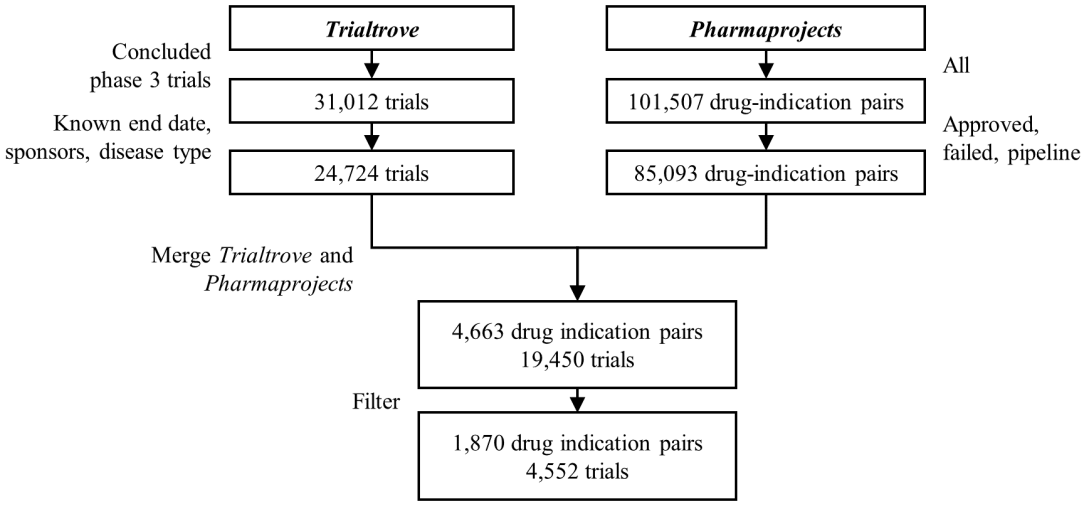


Figure B-2: Data cleaning for P3APP.

Table B.2: Examples of features extracted from Pharmaprojects and Trialtrove. After transforming multi-label parent features into binary child features, there were over 3,000 drug and trial categories in total. However, not all are useful for our analysis. For instance, trials rarely take place in Nepal, so the corresponding location feature rarely appears. Thus, this feature is unlikely to have meaningful associations with success. We remove these near zero variance factors. Also, we standardize continuous variables prior to all experiments.

	Examples	**Categories**
Drug Features		
Route	Inhaled; Injectable; Oral; Topical	4
Origin	Biological, protein, antibody; Biological, protein, recombinant; Chemical, synthetic	3
Medium	Capsule, hard; Capsule, soft; Powder; Solution; Suspension; Tablet	6
Biological target family	Cytokine/growth factor; Enzyme; Ion channel; Receptor; Transporter	5
Pharmacological target family	5 Hydroxytryptamine receptor antagonist; Angiogenesis inhibitor; Apoptosis stimulant; Cell cycle inhibitor; DNA inhibitor; DNA synthesis inhibitor; Growth factor receptor antagonist; Immunostimulant; Immunosuppressant; Ion channel antagonist; Protein kinase inhibitor	11
Drug-indication development status	True; false	2
Prior approval of drug for another indication	Approved; failed	2

Table B.2 (continued): Examples of features extracted from Pharmaprojects and Trialtrove.

	Examples	Categories
Trial Features		
Duration	Integer	1
Study design	Active comparator; Cross over; Dose response; Double blind/blinded; Efficacy; Multiple arm; Non-inferiority; Open label; Pharmacodynamics; Pharmacokinetics; Placebo control; Randomized; Safety; Single arm	14
Sponsor type	Academic; Cooperative Group; Government; Industry, all other pharma; Industry, top 20 Pharma	5
Therapeutic area	Autoimmune/Inflammation; Cardiovascular; CNS; Infectious Disease; Metabolic/Endocrinology; Oncology	6
Trial status	Completed; terminated	2
Trial outcomes	Completed, negative outcome or primary endpoint(s) not met; Completed, outcome indeterminate; Completed, positive outcome or primary endpoint(s) met; Terminated, business decision - other; Terminated, business decision - pipeline reprioritization; Terminated, lack of efficacy; Terminated, poor enrollment; Terminated, safety or adverse effects	8
Target accrual	Integer	1
Actual accrual	Integer	1
Locations	Argentina; Australia; Austria; Belgium; Brazil; Bulgaria; Canada; Chile; Czech Republic; Denmark; Europe; Finland; France; Germany; Hungary; India; Israel; Italy; Japan; Mexico; Netherlands; New Zealand; Peru; Poland; Romania; Russia; Slovakia; South Africa; South Korea; Spain; Sweden; Switzerland; Taiwan; Ukraine; United Kingdom; United States	36
Number of identified sites	Integer	1
Biomarker involvement	Biomarker/efficacy; Biomarker/toxicity; PGX - biomarker identification/evaluation; PGX - pathogen; PGX - patient preselection/stratification	5
Sponsor track record	Number of prior approved drug-indication pairs; Number of prior failed pairs; Total number of phase 1 trials sponsored; Number of phase 1 trials with positive results; Number of phase 1 trials with negative results; Number of completed phase 1 trials; Number of terminated phase 1 trials; Total number of phase 2 trials sponsored; Number of phase 2 trials with positive results; Number of phase 2 trials with negative results; Number of completed phase 2 trials; Number of terminated phase 2 trials; Total number of phase 3 trials sponsored; Number of phase 3 trials with positive results; Number of phase 3 trials with negative results; Number of completed phase 3 trials; Number of terminated phase 3 trials	17
Investigator experience	Refer to sponsor track record	17

B.2 Multiple Imputation

Multiple imputation is a principled missing data method that can provide valid statistical inferences when missingness is ignorable. It involves three steps: imputation, analysis, and pooling (see Fig. B-3).

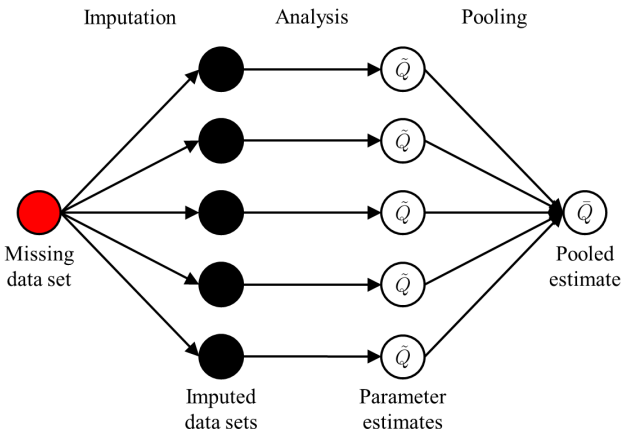


Figure B-3: Multiple imputation.

B.2.1 Imputation

Under MI, we draw multiple plausible values for each missing data point, thus creating multiple imputed datasets from one incomplete dataset. There are different strategies for multivariate multiple imputation. In this study, we focus on Fully Conditional Specification (FCS), specifically the Multivariate Imputation by Chained Equations (MICE) algorithm [73, 218]. In MICE, we first specify an imputation model for each incomplete variable in the form of conditional distributions (missing data conditioned on the observed data). The algorithm starts with simple random draws from the observed data and imputes the incomplete data in an iterative variable-by-variable fashion according to the specified variable models. Each iteration entails one cycle through all the incomplete variables (see Algo. B-2). The number of iterations should be set such that convergence is reached. This is typically checked by monitoring the means of imputed values and/or the values of regression coefficients and making sure they are stable over the iterations. In practice, a small number of iterations appears to

Algorithm B-1: Pseudo-code for Multivariate Imputation by Chained Equations.

Define Y as a $n \times p$ matrix where rows represent samples and columns represent variables
Data: Incomplete dataset $Y = (Y^{\text{obs}}, Y^{\text{mis}})$
Result: Imputed dataset $Y^T = (Y^{\text{obs}}, Y^{\text{mis},T})$ at iteration T
Define Y_j as the j^{th} feature column of Y where $Y_j = (Y_j^{\text{obs}}, Y_j^{\text{mis}})$
for $j \leftarrow 1$ **to** p **do**
    imputation model for incomplete variable $Y_j \leftarrow P(Y_j|Y_{-j}, \theta_j)$
    starting imputations $Y_j^{\text{mis},0} \leftarrow$ draws from Y_j^{obs}
Define $Y_{-j}^t = (Y_1^t, \ldots, Y_{j-1}^t, Y_{j+1}^{t-1}, \ldots, Y_p^{t-1})$ where Y_j^t is the j^{th} feature at iteration t
for $t \leftarrow 1$ **to** T **do**
    for $j \leftarrow 1$ **to** p **do**
        $\theta_j^t \leftarrow$ draw from posterior $P(\theta_j|Y_j^{\text{obs}}, Y_{-j}^t)$
        $Y_j^{\text{mis},t} \leftarrow$ draws from posterior predictive $P(Y_j^{\text{mis}}|Y_{-j}^t, \theta_j^t)$
return Y^T

be sufficient, from 10 to 20. Multiple imputed datasets can be generated by running MICE in parallel the desired number of times.

In this study, we specify linear regression models for incomplete continuous variables and logistic regression models for incomplete nominal variables. We monitor convergence by computing the mean/mode of the imputed values and making sure that they were stable over iterations. Twenty iterations appear to be sufficient.

B.2.2 Analysis

The analysis step after single imputation is straightforward: We apply any standard, complete-data statistical methods and end up with one set of results. In MI, we have multiple imputed compete datasets. After analyzing them individually using standard statistical procedures, we end up with multiple sets of results. These results will differ from each other since each dataset is imputed with different values. These differences represent the uncertainty due to the missing data. The pooling step describes how we can combine these sets of results into a single set.

B.2.3 Pooling

In this step, we pool the estimates obtained from multiple individual analyses using Rubin's rules to yield a single estimate [76]. Estimates that can be combined using

Rubin's rules include means, regression coefficients and probability predictions. Let Q be a column vector of the estimands of interest, $\tilde{Q}$ be its estimate, m be the number of imputed datasets, and $\tilde{Q}_l$ be the estimate of the l^{th} repeated analysis. The combined estimate is given by:

$$\bar{Q} = \frac{1}{m} \sum_{l=1}^{m} \tilde{Q}_l \tag{B.1}$$

B.3 Imputation Versus Listwise Deletion

We design an experiment to study the effects of imputation and verify that imputation indeed offers an improvement over complete cases analysis. First, we create a gold standard dataset by taking complete cases of the P2APP dataset (see Table 3.7). Next, we randomly split the gold standard dataset drug-indication pairs into a training set (70%) and a testing set (30%).

To simulate missingness present in the original dataset, we introduce missingness in the gold standard training and testing sets based on our MAR assumptions and the missingness patterns observed in the P2APP dataset. When making MAR, we ensure that the proportions of drugs and trials with fully observed features (i.e., complete cases) are consistent with those in the parent dataset (see Appendix B.3.1 for description).

We must be cautious relying on the MAR testing set for model validation. Results may not accurately capture whether a classifier has learned the true underlying relationship between the features and the outcome. To illustrate, suppose that drug-indication pairs have only one binary feature ("0" or "1") that is unrelated to approval/failure. Thus, no classifier can do better than random guessing (0.5 AUC). Now, assume that we have MAR in the dataset: failed pairs are more likely to have missing values due to the data collection process, unrelated to the binary feature. Suppose that we impute all the missing values with 1. Intuitively, we know that this is a poor imputation method because it distorts the feature distribution of failed pairs, and it reduces the variability in the data. However, this is seemingly a "good" method because it allows the AUC of a classifier on this imputed dataset to exceed 0.5. That is, we can identify a disproportionate number of failures by guessing all pairs with feature value 1 as failures. The classifier has learned a nonexistent relationship introduced by the imputations. By predicting all 1s as failures, the classifier is implicitly exploiting its MAR-ness.

Some may argue that it is acceptable to use missingness as a signal. Unfortunately, this is inappropriate in our case, because the MAR nature of the dataset on hand is

merely an artifact of data collection that would not be present during actual testing. MAR was introduced to the data due to the backfilling of information over time. This occurs due to a combination of reasons—some drug characteristics (e.g., mechanism of action) only become clear as the study progresses to higher phases; poor reporting practices. We believe that missingness in current test cases, e.g., drug-indication pairs currently in the pipeline, is more MCAR-like in nature because no backfilling has been performed. For example, immediately after phase 2 testing, pairs that go on to be approved are equally likely to have missing information as pairs that go on to be terminated. Clearly, missingness will not be a useful predictive factor. A classifier that relies heavily on the missingness in the dataset will fail miserably when put into production.

It is difficult to assess how good a classifier really is from the performance on a MAR testing set. Therefore, we create an additional testing set (the "MCAR testing set") in which we introduce missingness based on patterns observed in pipeline drug-indication pairs in the P2APP dataset (see Appendix B.3.1 for a description). Because the drugs were still in development at the time of snapshot of the databases, they are likely to be less affected by backfilling. Consequently, the AUC on the MCAR testing set will be more reflective of a classifier's real performance. We also use the gold standard testing sets for evaluation. These two testing sets serve as a control for the backfilling artifact in the data collection process. They can help to identify non-ideal imputation methods: poor imputation methods tend to distort the data distribution and undermine relationships between variables. This noise makes it more difficult for classifiers to learn the true underlying patterns in the data. These classifiers will perform poorly on the gold standard and MCAR testing sets. Returning to the above binary feature example, if we had tested the classifier on a gold standard testing set, we would realize that it did not learn any useful patterns. On the other hand, applying imputation methods that are capable of preserving the data distribution will make it easier for classifiers to capture useful relationships in the data. These classifiers will perform well on the gold standard and MCAR testing sets.

We have two training sets (gold standard and MAR) and three testing sets (gold

standard, MAR, and MCAR) (see Fig. B-4). We use five different missing data approaches, as described in Section 3.3.1, to generate multiple complete training sets from the MAR training set. Subsequently, we use each imputed training set to build six different predictive models (PLR, RF, NN, GBT, SVM, and C5.0) according to the methodology outlined in Section 3.3.2. We use ten-fold cross-validation to select the hyper-parameters for each model. In addition to the imputed MAR training sets, we use the gold standard training set to train gold standard classifiers: the models that would have been built if the data was complete. We impute the MAR and MCAR testing sets in a similar fashion as the training sets, and evaluate the AUC performance of all classifiers on the imputed and gold standard testing sets. We repeat the entire procedure of introducing MAR and MCAR in the dataset, imputing missingness, training models and validating performance 100 times for robustness. In addition to the AUC, we compute the biasness of the imputed values in the imputed training and testing sets with respect to their gold standard counterparts. This is a measure of accuracy of each imputation method. Finally, we use the results from the gold standard, MAR, and MCAR testing sets as basis to select an imputation method and machine-learning algorithm combination most suitable to the dataset on hand.

Table B.4 summarizes the results. Since the training and testing sets are fixed, using the same drug-indication pairs for all methods, direct comparison across different missing data techniques and machine-learning algorithms is possible. Each row corresponds to different missing data techniques used to process the training and testing sets in the experiments. Each column group corresponds to different types of missingness introduced in the testing sets. For all six machine-learning algorithms, we find that gold standard classifiers consistently outperform their complete cases analysis and imputation counterparts. This is logical because useful information is invariably lost when we intentionally introduce missingness in the datasets. In contrast, complete cases analysis often leads to inferior performance. The AUCs of classifiers trained on complete cases training sets are on average 0.04 less than those trained on imputed training sets. As expected, complete cases are ill suited for MAR data. This

supports our conjecture that the use of imputation has allowed predictive models to learn useful patterns that would otherwise be lost from discarding incomplete data.

When comparing across rows, we observe that the different imputation techniques are not equally effective. In terms of imputation quality, MI and mean/mode give the most inaccurate imputations while nearest neighbors recovers data best for both continuous and nominal variables (see Table B.3).

To better visualize each imputation method, Fig. B-5 plots the distributions of the trial feature of actual accrual, a continuous variable, in the gold standard, complete cases, and imputed MAR training sets of one iteration. It is evident that mean and median imputations have distorted the variable distribution, introducing previously absent peaks at the observed mean and median, respectively. In contrast, MI and nearest neighbors imputation managed to preserve the general shape of the variable distribution without introducing anomalous peaks.

We believe that the noise introduced by mean and median imputations have an adverse impact on a classifier's learning process. These effects may not be obvious from the AUC of the MAR testing sets. Indeed, for all six machine-learning algorithms, we observe that mean and median imputations give the highest AUCs for the MAR testing sets. However, the trend is reversed when we look at the gold standard and MCAR testing sets. Classifiers trained on mean or median imputation performed the worst of all imputation methods on these testing sets, implying that the noise introduced by the distortions must have hindered the machine-learning algorithms from fully capturing the underlying relationships in the data. It will therefore be prudent to avoid this imputation approach.

Overall, we find kNN imputation to be most suitable to the dataset. Note that the MI ($m = 10$)-RF and MI ($m = 10$)-C5.0 combinations yielded slightly better performances than kNN-RF. However, we excluded MI ($m = 10$) from consideration because the improvement is only marginal while the imputation and analysis processes are much more time consuming, since we have ten imputed datasets in MI ($m = 10$). Furthermore, the imputation method does not converge well (or at all) for smaller datasets. This poses an issue for the time series analysis in Section 3.4.3. In

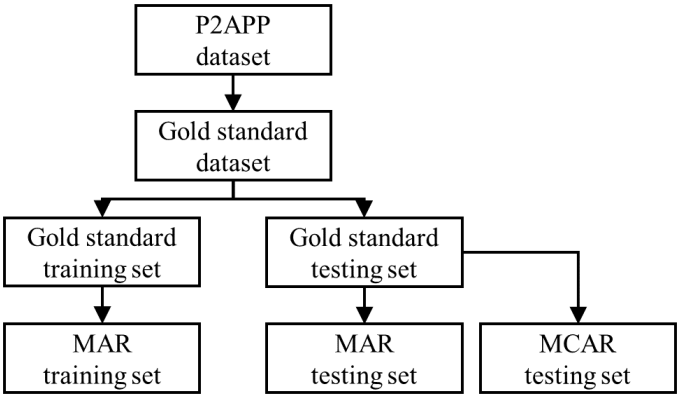


Figure B-4: Datasets created in the experiment.

contrast, kNN imputation is relatively straightforward to implement and more stable. It provides the least biased imputations among all missing data methods. More importantly, classifiers built on kNN-imputed training sets give the highest AUCs for the gold standard testing set for all machine-learning models explored. By preserving the original data distribution while filling in missing values, kNN imputation has allowed classifiers to learn underlying patterns more effectively. In particular, the combination of 5NN with RF gives the one of the highest gold standard (0.805) and MCAR (0.780) testing set AUCs. This may be attributed to the fact that RF is a nonlinear model, and thus it is able to better capture the complex interactions between the features and regulatory approval than PLR, a linear model. We focus on the 5NN-RF combination in our analyses, since it appears that this pair is most compatible with our datasets.

Table B.3: Biasness of imputations with respect to gold standard.

	MAR Training Set		**MAR Testing Set**		**MCAR Testing Set**	
	Bias[1]	**Wrongly Imputed**[2]	**Bias**[1]	**Wrongly Imputed**[2]	**Bias**[1]	**Wrongly Imputed**[2]
	%	%	%	%	%	%
Mean/mode	234.6	23.0	236.2	23.3	274.2	22.2
Median/mode	115.5	23.0	116.1	23.3	128.7	22.2
5NN	95.4	22.7	94.9	22.0	96.2	21.8
10NN	87.3	21.7	87.9	21.2	90.2	21.0
MI (m=1)	262.0	25.3	268.9	27.9	323.0	26.9
MI (m=10)	260.9	25.3	269.0	27.9	322.7	26.7

[1] Average percentage bias of imputed continuous variables. We first find the sum of the absolute percentage difference between imputed values that are continuous and their corresponding gold standard values (gold standard values as denominator), averaged over the total number of missing values that are continuous. Next, we take the mean over 100 iterations. [2] Percentage of nominal variables that were wrongly imputed. We first find the number of imputed categorical values that differ form their corresponding gold standard values, averaged over the total number of missing values that are categorical. Next, we take the mean over 100 iterations.

Table B.4: Performance of different missing data approaches. *Abbreviations*: avg, average; sd, standard deviation; 5%, 5th percentile; 50%, median; 95%, 95th percentile; m, number of imputations generated.

	Testing Set AUC														
	MAR					**MCAR**					**Gold Standard**				
	Avg	**Sd**	**5%**	**50%**	**95%**	**Avg**	**Sd**	**5%**	**50%**	**95%**	**Avg**	**Sd**	**5%**	**50%**	**95%**
PLR															
Gold Standard											0.810	0.028	0.761	0.808	0.853
Complete Cases											0.755	0.040	0.683	0.764	0.813
Mean/mode	0.786	0.028	0.746	0.785	0.829	0.751	0.029	0.702	0.753	0.794	0.778	0.031	0.729	0.779	0.823
Median/mode	0.786	0.028	0.745	0.786	0.829	0.751	0.029	0.704	0.753	0.794	0.778	0.031	0.728	0.779	0.824
5NN	0.763	0.032	0.716	0.762	0.814	0.757	0.032	0.707	0.758	0.805	0.786	0.032	0.738	0.787	0.834
10NN	0.774	0.030	0.730	0.773	0.821	0.757	0.032	0.695	0.756	0.802	0.787	0.032	0.739	0.791	0.835
MI ($m = 1$)	0.746	0.035	0.688	0.747	0.804	0.758	0.035	0.705	0.755	0.818	0.781	0.036	0.722	0.777	0.843
MI ($m = 10$)	0.755	0.030	0.705	0.757	0.801	0.766	0.032	0.719	0.764	0.815	0.782	0.031	0.729	0.782	0.831
RF															
Gold Standard											0.837	0.027	0.793	0.837	0.876
Complete Cases											0.764	0.048	0.685	0.772	0.830
Mean/mode	0.794	0.027	0.753	0.794	0.836	0.761	0.030	0.712	0.761	0.809	0.775	0.031	0.726	0.771	0.822
Median/mode	0.793	0.027	0.756	0.793	0.831	0.759	0.030	0.709	0.762	0.808	0.774	0.031	0.723	0.774	0.827
5NN	**0.782**	0.031	0.735	0.783	0.830	0.780	0.030	0.734	0.783	0.828	0.805	0.033	0.755	0.805	0.857
10NN	0.788	0.029	0.741	0.786	0.833	0.780	0.030	0.729	0.778	0.827	0.802	0.033	0.747	0.805	0.856
MI ($m = 1$)	0.774	0.028	0.732	0.777	0.825	0.782	0.031	0.737	0.779	0.845	0.797	0.033	0.748	0.795	0.853
MI ($m = 10$)	0.782	0.029	0.734	0.781	0.831	0.791	0.029	0.739	0.790	0.835	0.804	0.030	0.751	0.804	0.848
NN															
Gold Standard											0.800	0.032	0.754	0.799	0.849
Complete Cases											0.715	0.043	0.638	0.716	0.779
Mean/mode	0.789	0.030	0.736	0.790	0.835	0.766	0.037	0.709	0.766	0.819	0.790	0.037	0.739	0.789	0.848
Median/mode	0.788	0.030	0.742	0.788	0.835	0.766	0.034	0.711	0.766	0.818	0.789	0.036	0.740	0.792	0.849
5NN	0.776	0.030	0.730	0.776	0.821	0.771	0.035	0.715	0.774	0.823	0.794	0.032	0.743	0.798	0.842
10NN	0.784	0.034	0.724	0.785	0.842	0.773	0.039	0.702	0.776	0.831	0.797	0.036	0.737	0.798	0.851
MI ($m = 1$)	0.753	0.035	0.689	0.758	0.801	0.764	0.037	0.708	0.760	0.820	0.780	0.036	0.719	0.781	0.838
MI ($m = 10$)	0.774	0.028	0.729	0.774	0.816	0.784	0.031	0.725	0.789	0.827	0.795	0.030	0.750	0.795	0.838

Table B.4 (continued): Performance of different missing data approaches. *Abbreviations*: avg, average; sd, standard deviation; 5%, 5th percentile; 50%, median; 95%, 95th percentile; m, number of imputations generated.

	Testing Set AUC														
	MAR					**MCAR**					**Gold Standard**				
	Avg	**Sd**	**5%**	**50%**	**95%**	**Avg**	**Sd**	**5%**	**50%**	**95%**	**Avg**	**Sd**	**5%**	**50%**	**95%**
GBT															
Gold Standard											0.820	0.028	0.776	0.821	0.868
Complete Cases											0.746	0.050	0.659	0.756	0.816
Mean/mode	0.793	0.029	0.746	0.795	0.839	0.762	0.029	0.716	0.763	0.808	0.781	0.034	0.724	0.784	0.826
Median/mode	0.792	0.030	0.743	0.793	0.832	0.760	0.030	0.708	0.762	0.804	0.778	0.033	0.719	0.783	0.823
5NN	0.780	0.030	0.732	0.779	0.821	0.772	0.032	0.717	0.772	0.822	0.796	0.029	0.737	0.798	0.837
10NN	0.787	0.026	0.747	0.788	0.830	0.773	0.028	0.722	0.773	0.817	0.796	0.028	0.748	0.798	0.838
MI ($m = 1$)	0.763	0.031	0.714	0.762	0.812	0.773	0.031	0.727	0.768	0.820	0.796	0.031	0.747	0.796	0.847
MI ($m = 10$)	0.778	0.029	0.733	0.780	0.822	0.789	0.030	0.739	0.789	0.838	0.804	0.031	0.757	0.803	0.854
SVM															
Gold Standard											0.785	0.030	0.730	0.786	0.831
Complete Cases											0.733	0.053	0.650	0.741	0.795
Mean/mode	0.772	0.032	0.724	0.773	0.820	0.741	0.032	0.686	0.748	0.788	0.766	0.036	0.707	0.771	0.818
Median/mode	0.771	0.029	0.729	0.768	0.817	0.740	0.031	0.683	0.745	0.780	0.764	0.035	0.711	0.771	0.818
5NN	0.751	0.031	0.699	0.748	0.803	0.745	0.034	0.697	0.746	0.800	0.771	0.034	0.722	0.770	0.827
10NN	0.758	0.035	0.688	0.760	0.814	0.745	0.037	0.679	0.749	0.808	0.772	0.037	0.710	0.773	0.825
MI ($m = 1$)	0.731	0.035	0.676	0.732	0.788	0.741	0.033	0.684	0.745	0.790	0.760	0.035	0.696	0.762	0.813
MI ($m = 10$)	0.746	0.030	0.705	0.746	0.797	0.755	0.031	0.707	0.753	0.797	0.768	0.030	0.719	0.764	0.813
C5.0															
Gold Standard											0.800	0.033	0.758	0.800	0.844
Complete Cases											0.710	0.063	0.585	0.713	0.802
Mean/mode	0.764	0.033	0.711	0.768	0.810	0.734	0.032	0.675	0.737	0.777	0.758	0.039	0.698	0.762	0.816
Median/mode	0.764	0.038	0.708	0.761	0.825	0.735	0.041	0.676	0.736	0.797	0.754	0.043	0.679	0.751	0.823
5NN	0.756	0.036	0.703	0.753	0.816	0.749	0.038	0.695	0.745	0.805	0.772	0.038	0.715	0.772	0.843
10NN	0.759	0.035	0.696	0.762	0.807	0.747	0.037	0.687	0.749	0.799	0.770	0.035	0.710	0.771	0.822
MI ($m = 1$)	0.733	0.038	0.672	0.731	0.795	0.741	0.036	0.680	0.740	0.800	0.758	0.037	0.701	0.754	0.819
MI ($m = 10$)	0.786	0.030	0.738	0.786	0.836	0.793	0.031	0.738	0.797	0.842	0.807	0.031	0.756	0.808	0.857
MAR[1]	0.759	0.037	0.699	0.759	0.811	0.744	0.037	0.685	0.741	0.801	0.761	0.037	0.705	0.757	0.812

[1] For MAR, we leave the missingness as it is and rely on the decision tree algorithm to handle them internally.

B.3.1 Simulating Missingness

We simulate missingness in gold standard training and testing sets (see Table 3.7) based on our assumption of MAR and the missingness patterns observed in the P2APP dataset (see Table B.5 and Table B.6). For example, 36% of approved drugs in the P2APP dataset have some incomplete drug features. Accordingly, we randomly select 36% of approved drugs in the gold standard training set and introduce missingness in drug features according to the observed proportions to form the MAR training set, e.g., 6% of these drugs will have missing pharmacological target family values, 76% will have missing biological target family values, and so on. We repeat this process for failed drugs, completed trials, and terminated trials. At the end, we propagate the missing drug and trial features into the training set feature matrix, so that drug-indication pairs for the same drug have the same drug features missing in their feature vectors, and drug-indication pairs with the same trial have the same trial features missing. Conversely, when making the sets MAR, we ensure that the proportions of drugs and trials with fully observed features (i.e., complete cases) are consistent with that observed in the parent dataset, e.g., 64% of approved drugs in the MAR training set have complete drug features. We repeat this procedure for the gold standard testing set to form the MAR testing set.

We simulate MCAR in the gold standard testing set in a similar fashion to form the MCAR testing set. However, here we use unconditional missingness patterns observed in the pipeline dataset (see Table B.5 and Table B.6), instead of the known outcomes set where backfilling has occurred.

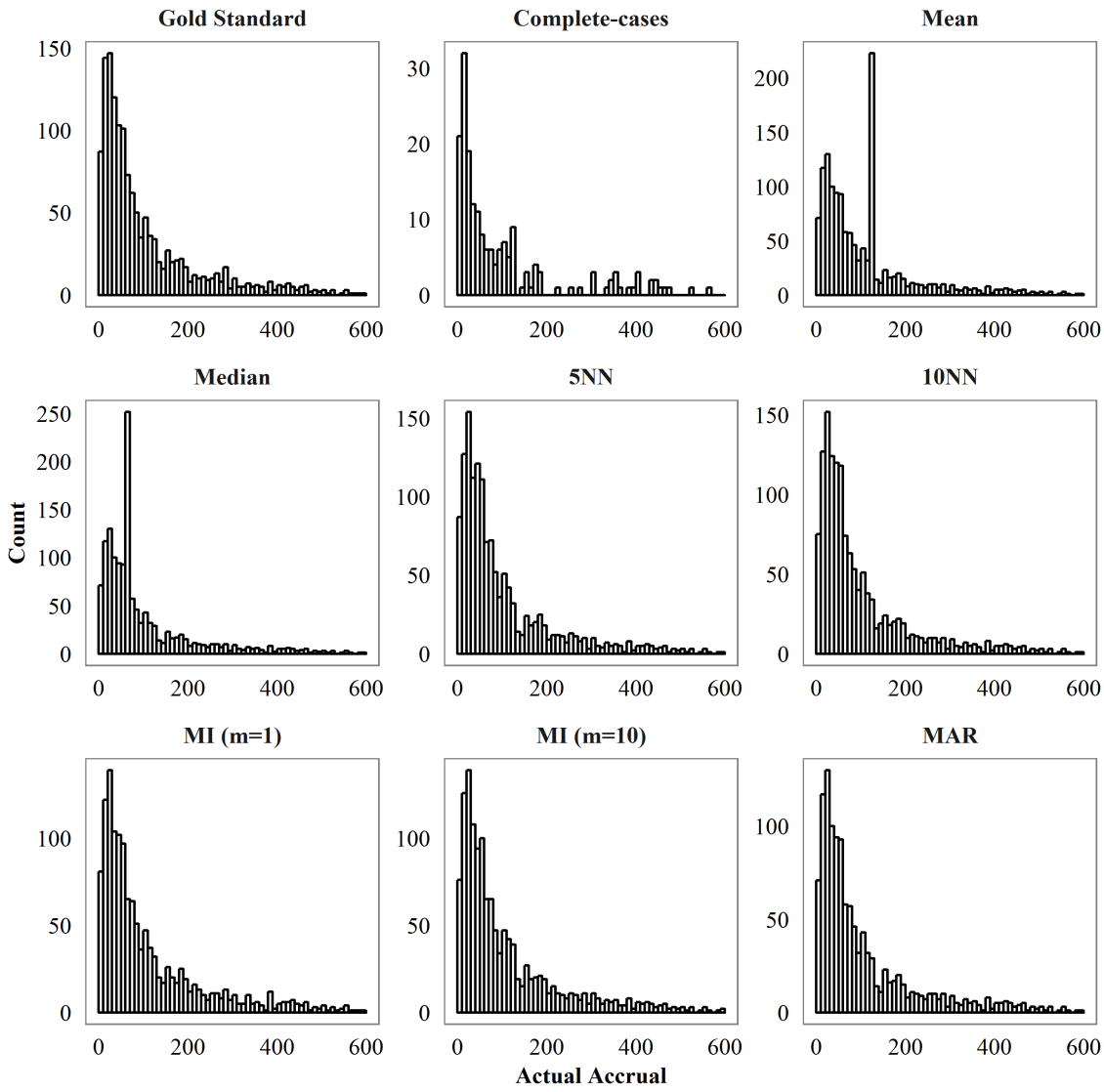


Figure B-5: Gold standard, complete cases, MAR, and imputed distributions of actual accrual in the training set of one of the iterations. The range of actual accrual goes up to 3,000. However, only a small number of samples go beyond 600. Thus, we truncated the histograms at 600 for better visualization. For MAR distribution, we ignored all missing values.

Table B.5: Breakdown of missingness in drug features in P2APP with respect to unique drugs.

	Missingness[1]		
	Known Outcomes		**Pipeline**
	Success	**Failure**	**Unconditional**
Complete Cases	0.64	0.29	0.46
Incomplete cases	0.36	0.71	0.54
Route	0.00	0.06	0.04
Pharmacological target family	0.06	0.10	0.17
Biological target family	0.76	0.45	0.63
Medium	0.43	0.86	0.69

[1] Feature missingness with respect to incomplete cases, e.g., 36% of success drugs have some incomplete drug features. 43% of these drugs have missing medium values.

Table B.6: Breakdown of missingness in trial features in P2APP with respect to unique trials.

	Missingness[1]		
	Known Outcomes		**Pipeline**
	Completion	**Termination**	**Unconditional**
Complete Cases	0.22	0.60	0.44
Incomplete cases	0.78	0.40	0.56
Number of identified sites	0.13	0.24	0.21
Actual accrual	0.13	0.54	0.18
Duration	0.37	0.13	0.24
Target accrual	0.54	0.21	0.37
Locations	0.02	0.04	0.02
Study design keywords	0.31	0.24	0.13
Trial outcomes	0.93	0.27	0.81

[1] Feature missingness with respect to incomplete cases.

B.4 Comparison with ANDI

The ANDI algorithm was proposed by DiMasi et al. [70] to predict regulatory approval for lead indications of cancer drugs after phase 2 testing. It is composed of a rubric of four factors to score anticancer agents (see Appendix B.4.1). The factors are based on pivotal trial characteristics and disease prevalence. Higher scores correspond to a higher probability of success. In this analysis, we apply ANDI on the oncology samples in the P2APP dataset, analyze its performance, and compare it with our 5NN-RF classifier in Appendix B.3.

First, we extract all cancer drugs from P2APP to form an oncology-only dataset. Since ANDI requires complete-cases, we drop all examples with missing values in any of the four ANDI factors (see Table B.7 for the resulting sample size). From this dataset, we draw a training set of 62 drugs with the same composition as that used by DiMasi et al. [70]: 40 failures and 22 successes. We set aside the remaining 319 drugs as a held-out testing set.

In replicating the ANDI experiment, we endeavored to follow the original proposed rubric as closely as possible. Unfortunately, two factors in the rubric are not in our dataset. We replace them with surrogate variables, and tune their cutoffs using the training set put aside earlier. The modified rubric is given in Table B.8. In order to apply ANDI, we have to identify the lead indication of each oncology drug and the pivotal phase 2 trial for that drug-indication pair. However, DiMasi et al. [70] did not provide clear instructions for identifying lead indications or pivotal trials. In this experiment, we apply heuristics which we felt were most logical. See Appendix B.4.1 for details on the proxy variables and heuristics used.

DiMasi et al. [70] reported an impressive 0.92 AUC for ANDI on a dataset of 62 drugs. However, this figure is based on in-sample/training-set testing, i.e., the algorithm was tested on the dataset on which the scoring rubric itself was derived. Such testing naturally yields excellent results because the four factors and their cutoffs were optimized for the algorithm to do well on the dataset. It is nearly impossible to judge whether an algorithm will generalize well without some form of testing on

held-out datasets. Unfortunately, such validation was not performed by DiMasi et al. [70]. Furthermore, ANDI was derived from a small sample, making it even more susceptible to overfitting.

For these reasons, it is very likely that the discriminative power of ANDI is actually much lower than that implied by the reported AUC of 0.92. Knowing these issues, we augment the ANDI experiment by including an out-of-sample model validation step, using the 319 drugs set aside as the testing set. This will allow us to determine ANDI's real performance more accurately.

The receiver operating characteristic curves of the original ANDI algorithm as reported in DiMasi et al. [70] and the modified ANDI on the oncology-only training and testing sets are shown in Fig. B-6. Similar to the original ANDI, our modified ANDI rubric demonstrates excellent performance on the training set with 0.94 AUC, 95% CI (0.89, 0.99). Unfortunately, this performance does not hold up on the testing set. The modified ANDI managed only 0.69 AUC on new, unseen samples. The large discrepancy between training and testing AUCs is indicative of overfitting. It is apparent that the patterns learned from the small training sample (n=62) do not generalize well, highlighting the importance of proper model validation. We believe the same holds for the original ANDI.

For a direct comparison with our classifiers, we apply the modified ANDI on oncology drugs in the gold standard testing sets in Appendix B.3. Fig. 3-7 summarizes the distributions of the results and compares 5NN-RF with the modified ANDI. On this testing set subsample, we find that our classifier achieves significantly higher AUC than the modified ANDI, an average improvement of 0.1 in AUC over 100 simulations. We believe that this gain can be attributed to a larger training set with a wider range of features, a nonlinear model that can capture the complex relationships in the data, and proper model validation methodology.

Lastly, we note that DiMasi et al. [70] applied complete-cases analysis in their study without any characterization of the missingness in their dataset. This is dangerous because complete-cases are appropriate only under strict MCAR conditions. Violation of these conditions will lead to biased estimates. Since data is rarely MCAR

Table B.7: Sample size of the oncology-only dataset (derived from P2APP).

	Counts				
	Drug-indication Pairs	**Phase 2 Trials**	**Unique Drugs**	**Unique Indications**	**Unique Phase 2 Trials**
Success	71	178	61	28	176
Failure	668	1,345	347	40	1,213
Total	739	1,523	381	40	1,368

Table B.8: Modified ANDI rubric in this study.

	Score		
	0	**1**	**2**
Trial outcomes[1]	Terminated, lack of efficacy; Completed, negative outcomes or primary endpoint(s) not met	Completed, outcome indeterminate	Terminated, early positive outcomes; Completed, positive outcomes or primary endpoint(s) met
Number of patients in pivotal phase 2 trial	≤ 37	38–49	≥ 50
US incidence[†]	$> 100{,}000$	10,000–100,000	$< 10{,}000$
Phase 2 duration (months)	> 44	21–44	< 21

[1] Surrogate variable.

in reality, it is unsurprising that the modified ANDI yields an inferior performance. In practice, this limits the applicability of ANDI to only samples with complete information. Given the scattershot nature of reporting in drug development, this makes ANDI less useful.

B.4.1 Modified ANDI

In replicating the ANDI experiment [70], we endeavored to follow the original proposed rubric as closely as possible (see Table B.9). Unfortunately, two factors in the rubric are not in our dataset: worldwide prevalence and activity. We replace them with surrogate variables, and tune their cutoffs using the training set placed aside earlier. The modified rubric is given in Table B.8. First, we use US incidence as a proxy for worldwide prevalence. This is because the latter figure is not known accurately for many of the oncology indications in our dataset, while the US incidence is much better documented and more accessible. (Sources include the American Cancer Society and the National Cancer Institute Surveillance, Epidemiology and End Re-

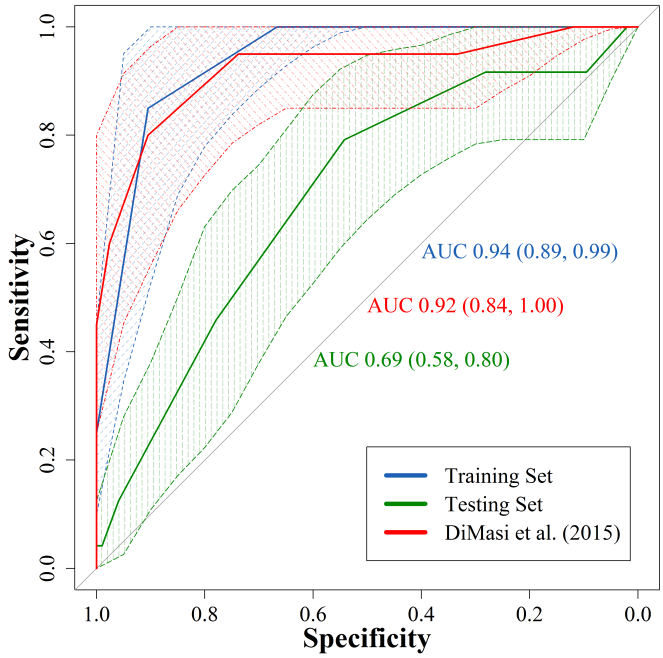


Figure B-6: Receiver operating characteristic curves of the original ANDI (as reported in [70]) and the modified ANDI on the oncology-only training and testing sets. We use bootstrapping to determine the 95% CI. We plot the receiver operating characteristic curve of the original ANDI in [70] (red) by using the ANDI scores breakdown provided in the study. The slight difference in the lower bound of the 95% CI between what we computed (0.84) and what DiMasi et al. [70] reported (0.81) may be accounted by randomness in the bootstraps. *Abbreviations*: ROC, receiver operating characteristic curve.

sults Program.) We determine the cutoffs in a manner similar to DiMasi et al. [70]: a larger incidence has lower scores while a smaller incidence has higher scores. Second, we use the trial outcome (i.e., the results of the trial) as a proxy for activity. We set the cutoffs similarly as in the original rubric: negative results have lower scores, while positive results have higher scores.

In order to apply ANDI, we have to identify the lead indication of each oncology drug and the pivotal phase 2 trial for that drug-indication pair. Unfortunately, DiMasi et al. [70] did not provide clear instructions for identifying lead indications or pivotal trials in the paper. Their attention was focused on what "they determined to be the lead cancer indication pursued," and they "identified what appeared to be the phase II trial that was most pivotal to the decision to proceed to large-scale phase III testing or to abandon the compound after phase II testing." It appears a fair amount of subjectivity is involved; there was no mention of any concrete criteria in the paper. This makes it difficult to replicate their study on other datasets. In this experiment, we apply heuristics which we felt were most logical. For drugs with multiple indications, we take the indication with the most phase 2 trials as the lead. We hypothesize that companies will invest in more trials for the designated lead indication. For drug-indication pairs with multiple phase 2 trials, we choose the trial with the largest accrual as the pivotal trial. This is logical, since trials with larger sample size have greater statistical power. They should hold greater weight in the decision of whether to proceed to phase 3 testing. In the event of ties, with the same number of trials or an identical accrual, we randomly select one of the candidates as the lead indication or pivotal trial.

Table B.9: Oncology ANDI proposed by DiMasi et al. [70].

	Score		
	0	**1**	**2**
Pivotal phase 2 trial activity	$< 3.0\%$ or negative randomized phase 2 trial	3.0–13.8%	$> 13.8\%$ or positive randomized phase 2 trial
Number of patients in pivotal phase 2 trial	≤ 37	38–49	≥ 50
Number of patients treated worldwide	$> 302{,}000$	50,000–302,000	$< 50{,}000$
Phase 2 duration (months)	> 44	21–44	< 21

B.5 Random Splitting Versus Temporal Ordering

We design an experiment to study the effects of any look-ahead bias introduced by splitting drug-indication pairs into training and testing sets randomly without considering the dates of development. First, we sample five-year rolling windows between 2004 and 2014 from the P2APP and P3APP datasets. In Section 3.4.3, we note that each window consists of a training set of drug-indication pairs whose outcomes become finalized within the window, and an out-of-sample, out-of-time testing set of drug-indication pairs that ended phase 2 or phase 3 testing, but are still in the pipeline with undetermined outcomes within the window. Here we disregard the temporal ordering—we aggregate the training and testing sets, and re-split them randomly before applying our machine-learning framework. To allow direct comparison with the time-series approach, we keep the new training and testing sample sizes same as those in Section 3.4.3. Table B.10 summarize the results.

We find that random splitting is indeed susceptible to overoptimistic performance (e.g., first four windows in P2APP). This may be attributed to the presence of future information in the training set, thus leading to look-ahead bias. However, we also observe over pessimistic results in some cases (e.g., last three windows in P3APP). This may occur when useful past information are set aside in the testing set. We believe that historical successes and failures contain valuable insights on the characteristics of high-potential candidates. Consider prediction for a phase 3 drug today. If we know that a drug with similar mechanism of action has been approved before, we should probably assign a higher chance of success to the pipeline drug under consideration. Conversely, if we see termination of drugs with similar mechanism of action in the past, we should lower our expectations for the pipeline drug as well. Under random allocation, the pipeline drug may be set aside in the testing set together with its historical counterpart. This prevents the model from learning from past experiences, which leads to over pessimistic performance.

The use of random splitting may be less than ideal due to the reasons noted above. It is prudent to adhere to the temporal ordering in the dataset when constructing

Table B.10: Comparison of classifiers trained on random splitting and temporal ordering. We use bootstrapping to determine the 95% CI for AUC.

	Sample Size		**Testing Set AUC (95% CI)**	
	Training Set	**Testing Set**	**Random Splitting**	**Temporal Ordering**
			P2APP	
2004-2008	1,361	551	0.750 (0.703, 0.797)	0.669 (0.614, 0.725)
2005-2009	1,562	591	0.764 (0.720, 0.808)	0.680 (0.625, 0.735)
2006-2010	1,764	636	0.748 (0.703, 0.794)	0.712 (0.668, 0.755)
2007-2011	1,969	598	0.768 (0.727, 0.809)	0.738 (0.698, 0.777)
2008-2012	2,082	597	0.750 (0.705, 0.795)	0.799 (0.760, 0.837)
2009-2013	2,212	517	0.781 (0.732, 0.829)	0.823 (0.779, 0.867)
2010-2014	2,289	380	0.795 (0.732, 0.858)	0.797 (0.718, 0.876)
			P3APP	
2004-2008	472	196	0.720 (0.650, 0.790)	0.769 (0.704, 0.834)
2005-2009	559	177	0.748 (0.675, 0.821)	0.724 (0.650, 0.798)
2006-2010	604	211	0.771 (0.707, 0.835)	0.738 (0.671, 0.805)
2007-2011	664	174	0.810 (0.743, 0.877)	0.806 (0.740, 0.871)
2008-2012	677	197	0.805 (0.744, 0.866)	0.827 (0.768, 0.886)
2009-2013	740	153	0.820 (0.754, 0.885)	0.868 (0.809, 0.927)
2010-2014	734	110	0.849 (0.772, 0.925)	0.876 (0.811, 0.941)

training and testing sets in order to obtain more realistic inferences.

B.6 Additional Results

Table B.11: Out-of-sample and out-of-time performance for P2APP. Comparison of the general and indication-group specific classifiers for selected indication groups. We use bootstrapping to determine the 95% CI for AUC. We exclude indication groups with too few samples.

	General			Indication-Group-Specific		
	Training Set	**Testing Set**	**Testing Set AUC (95% CI)**	**Training Set**	**Testing Set**	**Testing Set AUC (95% CI)**
			All			
2004-2008	1,361	551	0.669 (0.614, 0.725)			
2005-2009	1,562	591	0.680 (0.625, 0.735)			
2006-2010	1,764	636	0.712 (0.668, 0.755)			
2007-2011	1,969	598	0.738 (0.698, 0.777)			
2008-2012	2,082	597	0.799 (0.760, 0.837)			
2009-2013	2,212	517	0.823 (0.779, 0.867)			
2010-2014	2,289	380	0.797 (0.718, 0.876)			
			Alimentary			
2004-2008	1,361	86	0.494 (0.294, 0.694)	170	86	0.502 (0.310, 0.694)
2005-2009	1,562	93	0.613 (0.440, 0.785)	197	93	0.459 (0.287, 0.630)
2006-2010	1,764	80	0.589 (0.447, 0.731)	237	80	0.491 (0.321, 0.662)
2007-2011	1,969	77	0.707 (0.592, 0.821)	257	77	0.541 (0.396, 0.686)
2008-2012	2,082	67	0.802 (0.694, 0.909)	275	67	0.402 (0.252, 0.553)
2009-2013	2,212	58	0.834 (0.715, 0.954)	279	58	0.610 (0.441, 0.780)
2010-2014	2,289	39	0.670 (0.427, 0.913)	274	39	0.656 (0.414, 0.899)
			Cardiovascular			
2004-2008	1,361	39	0.515 (0.313, 0.717)	93	39	0.541 (0.310, 0.771)
2005-2009	1,562	38	0.307 (0.104, 0.509)	105	38	0.452 (0.230, 0.674)
2006-2010	1,764	46	0.613 (0.430, 0.795)	118	46	0.628 (0.449, 0.806)
2007-2011	1,969	37	0.634 (0.396, 0.872)	135	37	0.793 (0.644, 0.942)
2008-2012	2,082	42	0.640 (0.426, 0.853)	137	42	0.621 (0.425, 0.818)
2009-2013	2,212	35	0.360 (0.138, 0.582)	145	35	0.460 (0.272, 0.648)
2010-2014	2,289	19	0.529 (0.000, 1.000)	148	19	0.618 (0.000, 1.000)
			Anti-infective			
2004-2008	1,361	46	0.658 (0.502, 0.815)	124	46	0.645 (0.478, 0.812)
2005-2009	1,562	44	0.695 (0.525, 0.866)	146	44	0.707 (0.551, 0.863)
2006-2010	1,764	53	0.733 (0.568, 0.897)	161	53	0.708 (0.552, 0.864)
2007-2011	1,969	44	0.648 (0.479, 0.818)	171	44	0.592 (0.420, 0.763)
2008-2012	2,082	43	0.801 (0.666, 0.936)	165	43	0.815 (0.684, 0.945)
2009-2013	2,212	32	0.658 (0.454, 0.862)	169	32	0.649 (0.435, 0.864)
2010-2014	2,289	18	0.875 (0.708, 1.000)	167	18	0.750 (0.515, 0.985)
			Anti-cancer			
2004-2008	1,361	137	0.665 (0.528, 0.803)	456	137	0.683 (0.533, 0.833)
2005-2009	1,562	163	0.739 (0.618, 0.861)	494	163	0.635 (0.512, 0.758)
2006-2010	1,764	188	0.774 (0.702, 0.846)	546	188	0.726 (0.635, 0.816)
2007-2011	1,969	193	0.830 (0.773, 0.887)	618	193	0.746 (0.661, 0.831)
2008-2012	2,082	198	0.805 (0.717, 0.894)	682	198	0.760 (0.665, 0.855)
2009-2013	2,212	177	0.852 (0.783, 0.922)	736	177	0.786 (0.696, 0.876)
2010-2014	2,289	173	0.815 (0.691, 0.938)	791	173	0.803 (0.666, 0.940)

Table B.11 (continued): Out-of-sample and out-of-time performance for P2APP. Comparison of the general and indication-group specific classifiers for selected indication groups.

	General			Indication-Group-Specific		
	Training Set	**Testing Set**	**Testing Set AUC (95% CI)**	**Training Set**	**Testing Set**	**Testing Set AUC (95% CI)**
			Musculoskeletal			
2004-2008	1,361	35	0.765 (0.597, 0.933)	96	35	0.704 (0.512, 0.896)
2005-2009	1,562	38	0.716 (0.489, 0.944)	109	38	0.674 (0.472, 0.876)
2006-2010	1,764	35	0.634 (0.439, 0.830)	111	35	0.509 (0.276, 0.742)
2007-2011	1,969	37	0.737 (0.571, 0.903)	119	37	0.677 (0.493, 0.860)
2008-2012	2,082	36	0.884 (0.773, 0.995)	127	36	0.683 (0.462, 0.904)
2009-2013	2,212	26	0.792 (0.573, 1.000)	133	26	0.667 (0.429, 0.904)
2010-2014	2,289	19	0.882 (0.724, 1.000)	128	19	0.882 (0.706, 1.000)
			Neurological			
2004-2008	1,361	122	0.688 (0.572, 0.803)	211	122	0.768 (0.676, 0.859)
2005-2009	1,562	119	0.612 (0.471, 0.753)	271	119	0.625 (0.501, 0.748)
2006-2010	1,764	125	0.656 (0.532, 0.779)	334	125	0.673 (0.560, 0.787)
2007-2011	1,969	105	0.701 (0.580, 0.822)	375	105	0.649 (0.522, 0.776)
2008-2012	2,082	114	0.806 (0.707, 0.904)	382	114	0.695 (0.586, 0.804)
2009-2013	2,212	87	0.938 (0.857, 1.000)	417	87	0.718 (0.558, 0.879)
2010-2014	2,289	55	0.984 (0.952, 1.000)	408	55	0.860 (0.721, 0.999)
			Respiratory			
2004-2008	1,361	34	0.673 (0.418, 0.927)	89	34	0.833 (0.650, 1.000)
2005-2009	1,562	42	0.842 (0.722, 0.962)	104	42	0.825 (0.670, 0.979)
2006-2010	1,764	49	0.797 (0.663, 0.931)	125	49	0.801 (0.644, 0.959)
2007-2011	1,969	36	0.694 (0.513, 0.875)	143	36	0.519 (0.323, 0.715)
2008-2012	2,082	43	0.751 (0.604, 0.899)	149	43	0.692 (0.520, 0.865)
2009-2013	2,212	37	0.827 (0.694, 0.961)	154	37	0.876 (0.764, 0.987)
2010-2014	2,289	23	0.724 (0.365, 1.000)	160	23	0.842 (0.679, 1.000)
			Rare Diseases			
2004-2008	1,361	69	0.664 (0.517, 0.811)	212	69	0.521 (0.349, 0.693)
2005-2009	1,562	81	0.627 (0.471, 0.782)	231	81	0.528 (0.368, 0.687)
2006-2010	1,764	108	0.774 (0.666, 0.881)	257	108	0.691 (0.546, 0.836)
2007-2011	1,969	101	0.786 (0.698, 0.874)	303	101	0.680 (0.547, 0.812)
2008-2012	2,082	112	0.787 (0.696, 0.879)	329	112	0.600 (0.469, 0.731)
2009-2013	2,212	90	0.803 (0.702, 0.903)	358	90	0.730 (0.626, 0.834)
2010-2014	2,289	89	0.793 (0.621, 0.965)	391	89	0.779 (0.626, 0.932)

Table B.12: Out-of-sample and out-of-time performance for P3APP. Comparison of the general and indication-group specific classifiers for selected indication groups. We use bootstrapping to determine the 95% CI for AUC. We exclude indication groups with too few samples.

	General			Indication-Group-Specific		
	Training Set	**Testing Set**	**Testing Set AUC (95% CI)**	**Training Set**	**Testing Set**	**Testing Set AUC (95% CI)**
			All			
2004-2008	472	196	0.769 (0.704, 0.834)			
2005-2009	559	177	0.724 (0.650, 0.798)			
2006-2010	604	211	0.738 (0.671, 0.805)			
2007-2011	664	174	0.806 (0.740, 0.871)			
2008-2012	677	197	0.827 (0.768, 0.886)			
2009-2013	740	153	0.868 (0.809, 0.927)			
2010-2014	734	110	0.876 (0.811, 0.941)			
			Alimentary			
2004-2008	472	65	0.826 (0.651, 1.000)	25	65	0.889 (0.756, 1.000)
2005-2009	559	75	0.683 (0.324, 1.000)	17	75	0.650 (0.331, 0.969)
2006-2010	604	80	0.672 (0.428, 0.915)	30	80	0.651 (0.429, 0.872)
2007-2011	664	91	0.911 (0.786, 1.000)	28	91	0.800 (0.630, 0.970)
2008-2012	677	97	0.786 (0.572, 1.000)	24	97	0.700 (0.469, 0.931)
2009-2013	740	107	0.607 (0.149, 1.000)	18	107	0.786 (0.570, 1.000)
2010-2014	734	99	0.944 (0.850, 1.000)	19	99	0.733 (0.492, 0.975)
			Anti-cancer			
2004-2008	472	95	0.773 (0.618, 0.928)	34	95	0.684 (0.495, 0.874)
2005-2009	559	107	0.740 (0.543, 0.936)	28	107	0.568 (0.345, 0.791)
2006-2010	604	110	0.754 (0.599, 0.910)	50	110	0.630 (0.452, 0.809)
2007-2011	664	132	0.587 (0.333, 0.842)	24	132	0.392 (0.132, 0.651)
2008-2012	677	134	0.793 (0.549, 1.000)	40	134	0.668 (0.457, 0.879)
2009-2013	740	151	0.800 (0.480, 1.000)	29	151	0.775 (0.528, 1.000)
2010-2014	734	153	0.943 (0.842, 1.000)	26	153	0.852 (0.558, 1.000)
			Neurological			
2004-2008	472	118	0.851 (0.753, 0.949)	59	118	0.837 (0.735, 0.939)
2005-2009	559	151	0.782 (0.646, 0.918)	45	151	0.784 (0.649, 0.919)
2006-2010	604	169	0.732 (0.593, 0.871)	52	169	0.759 (0.629, 0.890)
2007-2011	664	180	0.706 (0.532, 0.880)	40	180	0.698 (0.529, 0.867)
2008-2012	677	178	0.765 (0.604, 0.926)	41	178	0.743 (0.586, 0.900)
2009-2013	740	185	0.827 (0.681, 0.973)	31	185	0.805 (0.641, 0.968)
2010-2014	734	166	0.779 (0.567, 0.990)	27	166	0.900 (0.782, 1.000)
			Rare Diseases			
2004-2008	472	54	0.711 (0.465, 0.957)	22	54	0.620 (0.364, 0.876)
2005-2009	559	60	0.735 (0.517, 0.952)	23	60	0.606 (0.360, 0.852)
2006-2010	604	66	0.888 (0.747, 1.000)	24	66	0.825 (0.645, 1.000)
2007-2011	664	72	0.838 (0.652, 1.000)	22	72	0.735 (0.520, 0.950)
2008-2012	677	76	0.893 (0.780, 1.000)	34	76	0.700 (0.523, 0.877)
2009-2013	740	94	0.962 (0.899, 1.000)	28	94	0.932 (0.840, 1.000)
2010-2014	734	109	0.908 (0.766, 1.000)	18	109	0.985 (0.942, 1.000)

Appendix C

Supplement to Chapter 4

C.1 Asymptotics for Superiority-by-Margin Testing

The constraint is:

$$\hat{p}_1 = \theta \hat{p}_0 \tag{C.1}$$

where $\hat{p}_1$ and $\hat{p}_0$ are the constrained maximum likelihood estimates of P_1 and P_0, respectively, under the null hypothesis.

The closed-form solution is given by:

$$\hat{p}_0 = \frac{-B - \sqrt{B^2 - 4AC}}{2A} \tag{C.2}$$

$$A = (r+1)\theta n_0 \quad , \quad B = -(\theta r n_0 + c_1 + n_0 + \theta c_0) \quad , \quad C = c_1 + c_0$$

The asymptotic approximation is:

$$\tilde{p}_0 = \frac{-B - \sqrt{B^2 - 4AC}}{2A} \quad , \quad \tilde{p}_1 = \theta \tilde{p}_0 \tag{C.3}$$

$$A = (r+1)\theta \quad , \quad B = -(\theta r + r P_1 + 1 + \theta P_0) \quad , \quad C = r P_1 + P_0$$

C.2 Parameter Estimation for the SIRDC-SD Model

Let D_t and d_t be the cumulative and daily number of deaths from data at time t, respectively. Let $\hat{D}_t$ and $hatd_t$ denote the estimated values of D_t and d_t, respectively. We use the following optimization program to estimate the parameters of the model:

$$\min_{\beta_0,\beta^*,\lambda,I_0,\eta} \quad \ln\left(\sum_t (D_t - \hat{D}_t)^2\right) + \ln\left(\sum_t (d_t - \hat{d}_t)^2\right) \tag{C.4}$$

subject to:

$$I_0 < N \ , \tag{C.5}$$

$$R_0 = \eta I_0 \ , \tag{C.6}$$

$$S_0 = N - R_0 - I_0 \ , \tag{C.7}$$

$$\beta_0 > \beta^* \ . \tag{C.8}$$

Our loss function is given by Eq. (C.4), which says that we minimize the sum of 1) the natural logarithm of the sum of squared errors for the cumulative deaths, and 2) the natural logarithm of the sum of squared errors for the daily deaths. The minimization program is subjected to the four constraints. Eq. (C.5) says that the initial number of infected must be less than the entire population. Eq. (C.6) imposes that the number of initial resolving cases must be less than the number of initial infected cases. Eq. (C.7) states that the conservation of population must hold at the start of the simulation and Eq. (C.8) constrains the initial contact rate to be greater than the final contact rate. We set γ, δ, and θ to 0.2, 0.008, and 0.1, respectively [117]. We solve the optimization program using the constrained Trust-Region algorithm as implemented in the SciPy Optimize package.

C.3 SIRDCV Model

We let $\bar{V}$ and ϵ be the number of persons vaccinated at every time step and the effectiveness of the vaccine, respectively. Effectiveness is defined as the performance of the vaccine under real-world conditions in a general population whereas efficacy is defined as the ability to protect against a virus under ideal conditions in a homogeneous population. The former is usually is less than the latter due to several reasons, e.g., improper storage of vaccines leading to loss of potency and non-compliance with the vaccine dosing schedule. For simplicity, we assume that the effectiveness of the vaccine in the epidemiological model is identical to the efficacy of the vaccine in the clinical trials.

$$\frac{dS_t}{dt} = -\frac{\beta(t)S_tI_t}{N} - \bar{V} \tag{C.9}$$

$$\frac{dI_t}{dt} = \frac{\beta(t)(S_t + V_t^{nr})I_t}{N} - \gamma I_t \tag{C.10}$$

$$\frac{dV_t^{nr}}{dt} = (1-\epsilon)\bar{V} - \frac{\beta(t)V_t^{nr}I_t}{N} \tag{C.11}$$

$$\frac{dV_t^{r}}{dt} = \epsilon\bar{V} \tag{C.12}$$

$$\frac{dR_t}{dt} = \gamma I_t - \theta R_t \tag{C.13}$$

$$\frac{dD_t}{dt} = \delta\theta R_t \tag{C.14}$$

$$\frac{dC_t}{dt} = (1-\delta)\theta R_t \tag{C.15}$$

where V_t^r and V_t^{nr} represent the stock of people who are inoculated, and respond (r) and do not respond (nr) to the vaccine, respectively. Eq. (4.13) has been modified to remove vaccinated persons at every time step in Eq. (C.9). We also modify Eq. (4.14) to allow people who are vaccinated but do not respond to the inoculation to be infected in Eq. (C.10). Eq. (C.11) and Eq. (C.12) keep track of the stock of people who are vaccinated. With this specification, the virus is allowed to spread even when the entire population is vaccinated because not everyone will respond to the mass inoculation.

C.4 Evolution of the Epidemic

We model three different scenarios regarding the evolution of the epidemic after lockdown is relaxed. We define β_{ss} as $max(0.22, \beta(T_v))$, where $\beta(T_v)$ is the value of β when the lockdown is released.

C.4.1 Status Quo

For the "status quo" scenario, we use the estimated dynamic $\beta(t)$ for our forecasts.

C.4.2 Ramp

For the "ramp" scenario, we model $\beta(t)$ according to:

$$\beta'(t) = \begin{cases} \beta(t) & \forall\, t < T_v \\ \beta(T_v) + \frac{\beta_{ss} - \beta(T_v)}{90} t & \forall\, T_v \leq t \leq (T_v + 90) \\ \beta_{ss} & \text{otherwise} \end{cases} \tag{C.16}$$

C.4.3 Behavioral

The "behavioral" scenario is modeled by making the percentage change in contact rate parameter negatively proportionate to the change in the observed death rate over an interval of t_o. That is,

$$\frac{1}{\beta} \frac{d\beta}{d(\frac{\Delta D}{N})} = -k \tag{C.17}$$

The contact rate parameter in this case is defined by:

$$\beta'(t) = \begin{cases} \beta(t) & \forall\, t < T_v \\ e^{c - k \frac{D_t - D_{t-t_o}}{N}} & \text{otherwise} \end{cases} \tag{C.18}$$

C.5 Financial Costs of Vaccine Efficacy Studies

There are many sources of costs involved in a clinical trial, e.g., patient recruitment and retention, medical and administrative staff, clinical procedures and central laboratory, site management, and data collection and analysis. For a back-of-the-envelope calculation, we assume that the cost per subject in a phase 3 vaccine efficacy trial is around US$5,000. This suggests a cost of US$150M for a study with 30,000 subjects, close to that estimated for rotavirus vaccines [219] in one of the very few studies that estimate the cost of vaccine development [220]. The figure is very high as compared to the median expense of a phase 3 trial for novel therapeutic agents, estimated to be US$19M [221]. However, this is not surprising because vaccine efficacy studies are notorious for being costly due to the large sample sizes and lengthy follow-up durations.

If we assume that challenge studies have a cost per subject that is ten times higher, i.e., US$50,000 per volunteer, the estimated cost of an HCT is approximately US$37.5M, where we have assumed a cost of US$5,000 per subject for the follow-up single-arm safety study comprising of 5,000 subjects. This makes up just 25% of the cost of an RCT with 30,000 subjects. Assuming a mortality rate of 1% in HCT—which is an overestimation given that the case fatality rate of adults aged 18–25 years has been estimated to be approximately 0.2% [107]—and a value of statistical life of $10M, the expected cost of liability compensation is $25M. This brings the expected cost of an HCT to $62.5M, approximately 40% of the cost of a traditional phase 3 vaccine efficacy field trial.

C.6 Additional Results

Table C.1: Expected number of incremental infections and deaths avoided in the U.S. under different trial designs, vaccine efficacies, and epidemiological scenarios, assuming trials start on August 1, 2020, superiority testing, and 1M doses of a vaccine per day are available after licensure, compared to the baseline case where no vaccine is ever approved.

	Vaccine Efficacy (%)			
	30		**50**	
	$\mathbb{E}[\Delta\text{Infections}]$	$\mathbb{E}[\Delta\text{Deaths}]$	$\mathbb{E}[\Delta\text{Infections}]$	$\mathbb{E}[\Delta\text{Deaths}]$
Status Quo				
RCT	2,506	20	8,116	64
ORCT	3,654	29	11,947	95
ARCT	6,248	49	22,261	177
HCT (30-day set-up)	90,472	722	106,202	848
HCT (60-day set-up)	71,223	568	83,467	666
HCT (90-day set-up)	56,263	448	65,857	525
HCT (120-day set-up)	44,556	355	52,122	415
Behavioral				
RCT	224,835	1,736	264,810	2,056
ORCT	705,881	5,591	925,920	7,344
ARCT	1,502,846	11,959	2,051,223	16,346
HCT (30-day set-up)	2,209,905	17,618	2,695,582	21,502
HCT (60-day set-up)	1,611,969	12,834	1,951,336	15,548
HCT (90-day set-up)	1,190,836	9,465	1,429,078	11,370
HCT (120-day set-up)	894,225	7,092	1,065,008	8,457
Ramp				
RCT	756,692	5,764	845,731	6,477
ORCT	1,825,095	14,344	2,656,479	20,964
ARCT	3,594,521	28,466	5,131,954	40,766
HCT (30-day set-up)	5,526,735	43,930	6,565,535	52,235
HCT (60-day set-up)	4,282,314	33,975	5,086,688	40,404
HCT (90-day set-up)	3,311,292	26,206	3,926,171	31,120
HCT (120-day set-up)	2,564,645	20,233	3,031,075	23,959

Table C.1 (continued): Expected number of incremental infections and deaths avoided in the U.S. under different trial designs, vaccine efficacies, and epidemiological scenarios, assuming trials start on August 1, 2020, superiority testing, and 1M doses of a vaccine per day are available after licensure, compared to the baseline case in which no vaccine is ever approved.

	Vaccine Efficacy (%)			
	70		**90**	
	$\mathbb{E}[\Delta\text{Infections}]$	$\mathbb{E}[\Delta\text{Deaths}]$	$\mathbb{E}[\Delta\text{Infections}]$	$\mathbb{E}[\Delta\text{Deaths}]$
Status Quo				
RCT	14,162	112	16,506	130
ORCT	25,167	200	38,663	308
ARCT	49,396	393	63,896	508
HCT (30-day set-up)	114,847	918	120,945	966
HCT (60-day set-up)	90,167	720	94,885	758
HCT (90-day set-up)	71,088	567	74,766	597
HCT (120-day set-up)	56,235	449	59,123	471
Behavioral				
RCT	289,168	2,251	306,050	2,386
ORCT	1,007,301	7,995	1,065,183	8,459
ARCT	2,269,753	18,094	2,423,075	19,321
HCT (30-day set-up)	2,982,094	23,794	3,189,157	25,451
HCT (60-day set-up)	2,150,531	17,142	2,294,765	18,295
HCT (90-day set-up)	1,566,872	12,473	1,666,446	13,269
HCT (120-day set-up)	1,161,296	9,228	1,230,321	9,780
Ramp				
RCT	899,765	6,909	937,666	7,212
ORCT	2,890,096	22,832	3,047,293	24,089
ARCT	5,768,903	45,861	6,091,608	48,443
HCT (30-day set-up)	7,130,975	56,759	7,523,068	59,896
HCT (60-day set-up)	5,528,656	43,941	5,837,268	46,409
HCT (90-day set-up)	4,265,392	33,834	4,503,392	35,738
HCT (120-day set-up)	3,288,349	26,018	3,469,234	27,465

Table C.2: Expected number of incremental infections and deaths avoided in the U.S. under different trial designs, vaccine efficacies, and epidemiological scenarios, assuming trials start on August 1, 2020, superiority testing, and infinite doses of a vaccine per day are available after licensure, compared to the baseline case in which no vaccine is ever approved.

	Vaccine Efficacy (%)			
	30		**50**	
	$\mathbb{E}[\Delta\text{Infections}]$	$\mathbb{E}[\Delta\text{Deaths}]$	$\mathbb{E}[\Delta\text{Infections}]$	$\mathbb{E}[\Delta\text{Deaths}]$
Status Quo				
RCT	4,343	35	12,691	101
ORCT	6,190	50	18,462	147
ARCT	10,655	84	34,672	276
HCT (30-day set-up)	157,044	1,255	168,612	1,347
HCT (60-day set-up)	122,531	978	131,429	1,049
HCT (90-day set-up)	96,093	767	102,986	822
HCT (120-day set-up)	75,691	604	81,068	647
Behavioral				
RCT	401,196	3,147	422,644	3,318
ORCT	1,284,033	10,217	1,542,261	12,276
ARCT	2,957,024	23,592	3,683,384	29,403
HCT (30-day set-up)	4,466,352	35,669	4,884,898	39,016
HCT (60-day set-up)	3,196,408	25,510	3,494,817	27,895
HCT (90-day set-up)	2,291,219	18,268	2,500,498	19,941
HCT (120-day set-up)	1,659,356	13,214	1,805,003	14,377
Ramp				
RCT	1,174,517	9,107	1,229,484	9,547
ORCT	3,172,803	25,126	4,242,057	33,649
ARCT	6,347,189	50,488	8,191,884	65,245
HCT (30-day set-up)	9,669,217	77,070	10,366,266	82,641
HCT (60-day set-up)	7,564,062	60,228	8,126,045	64,719
HCT (90-day set-up)	5,860,161	46,598	6,304,440	50,146
HCT (120-day set-up)	4,512,448	35,815	4,857,257	38,569

Table C.2 (continued): Expected number of incremental infections and deaths avoided in the U.S. under different trial designs, vaccine efficacies, and epidemiological scenarios, assuming trials start on August 1, 2020, superiority testing, and infinite doses of a vaccine per day are available after licensure, compared to the baseline case in which no vaccine is ever approved.

	Vaccine Efficacy (%)			
	70		**90**	
	$\mathbb{E}[\Delta\text{Infections}]$	$\mathbb{E}[\Delta\text{Deaths}]$	$\mathbb{E}[\Delta\text{Infections}]$	$\mathbb{E}[\Delta\text{Deaths}]$
Status Quo				
RCT	20,900	165	23,426	185
ORCT	36,872	294	54,672	436
ARCT	72,976	581	90,989	725
HCT (30-day set-up)	172,598	1,380	174,917	1,398
HCT (60-day set-up)	134,478	1,075	136,254	1,088
HCT (90-day set-up)	105,338	841	106,709	852
HCT (120-day set-up)	82,896	662	83,965	670
Behavioral				
RCT	432,235	3,396	437,725	3,439
ORCT	1,587,101	12,634	1,613,158	12,843
ARCT	3,813,885	30,447	3,881,898	30,991
HCT (30-day set-up)	5,039,465	40,253	5,128,348	40,964
HCT (60-day set-up)	3,605,985	28,786	3,670,305	29,300
HCT (90-day set-up)	2,578,527	20,566	2,623,871	20,928
HCT (120-day set-up)	1,858,914	14,809	1,890,330	15,060
Ramp				
RCT	1,255,157	9,752	1,270,085	9,871
ORCT	4,362,661	34,612	4,422,914	35,094
ARCT	8,662,725	69,012	8,776,472	69,922
HCT (30-day set-up)	10,597,019	84,487	10,728,517	85,539
HCT (60-day set-up)	8,315,537	66,236	8,423,946	67,103
HCT (90-day set-up)	6,456,348	51,362	6,543,545	52,059
HCT (120-day set-up)	4,976,272	39,521	5,044,819	40,070

Table C.3: Expected number of incremental infections and deaths avoided in the U.S. under different trial designs, vaccine efficacies, and epidemiological scenarios, assuming trials start on August 1, 2020, superiority-by-margin testing at 30%, and 1M doses of a vaccine per day are available after licensure, compared to the baseline case where no vaccine is ever approved.

	Vaccine Efficacy (%)			
	30		**50**	
	$\mathbb{E}[\Delta\text{Infections}]$	$\mathbb{E}[\Delta\text{Deaths}]$	$\mathbb{E}[\Delta\text{Infections}]$	$\mathbb{E}[\Delta\text{Deaths}]$
Status Quo				
RCT	263	2	2,618	21
ORCT	999	8	4,251	34
ARCT	369	2	3,735	29
HCT (30-day set-up)	2,139	17	99,609	795
HCT (60-day set-up)	1,648	13	78,283	625
HCT (90-day set-up)	1,267	10	61,765	492
HCT (120-day set-up)	969	8	48,882	389
Behavioral				
RCT	2,252	18	264,786	2,056
ORCT	18,752	149	746,378	5,915
ARCT	26,078	207	1,635,970	13,024
HCT (30-day set-up)	56,145	448	2,528,441	20,169
HCT (60-day set-up)	40,908	326	1,830,340	14,584
HCT (90-day set-up)	30,177	240	1,340,463	10,665
HCT (120-day set-up)	22,619	180	998,966	7,933
Ramp				
RCT	11,528	88	845,618	6,476
ORCT	56,093	442	1,893,630	14,903
ARCT	74,754	590	3,823,126	30,295
HCT (30-day set-up)	140,662	1,118	6,158,447	48,996
HCT (60-day set-up)	108,952	865	4,771,293	37,899
HCT (90-day set-up)	84,209	667	3,682,731	29,190
HCT (120-day set-up)	65,184	515	2,843,132	22,473

Table C.3 (continued): Expected number of incremental infections and deaths avoided in the U.S. under different trial designs, vaccine efficacies, and epidemiological scenarios, assuming trials start on August 1, 2020, superiority-by-margin testing at 30%, and 1M doses of a vaccine per day are available after licensure, compared to the baseline case in which no vaccine is ever approved.

	Vaccine Efficacy (%)			
	70		**90**	
	$\mathbb{E}[\Delta\text{Infections}]$	$\mathbb{E}[\Delta\text{Deaths}]$	$\mathbb{E}[\Delta\text{Infections}]$	$\mathbb{E}[\Delta\text{Deaths}]$
Status Quo				
RCT	10,240	81	16,273	129
ORCT	16,049	128	34,943	278
ARCT	20,883	166	50,277	400
HCT (30-day set-up)	114,847	918	120,945	966
HCT (60-day set-up)	90,167	720	94,885	758
HCT (90-day set-up)	71,088	567	74,766	597
HCT (120-day set-up)	56,235	449	59,123	471
Behavioral				
RCT	289,168	2,251	306,050	2,386
ORCT	1,007,287	7,995	1,065,183	8,459
ARCT	2,266,473	18,068	2,423,075	19,321
HCT (30-day set-up)	2,982,094	23,794	3,189,157	25,451
HCT (60-day set-up)	2,150,531	17,142	2,294,765	18,295
HCT (90-day set-up)	1,566,872	12,473	1,666,446	13,269
HCT (120-day set-up)	1,161,296	9,228	1,230,321	9,780
Ramp				
RCT	899,765	6,909	937,666	7,212
ORCT	2,887,058	22,808	3,047,293	24,089
ARCT	5,629,215	44,744	6,091,608	48,443
HCT (30-day set-up)	7,130,975	56,759	7,523,068	59,896
HCT (60-day set-up)	5,528,656	43,941	5,837,268	46,409
HCT (90-day set-up)	4,265,392	33,834	4,503,392	35,738
HCT (120-day set-up)	3,288,349	26,018	3,469,234	27,465

Table C.4: Expected number of incremental infections and deaths avoided in the U.S. under different trial designs, vaccine efficacies, and epidemiological scenarios, assuming trials start on August 1, 2020, superiority-by-margin testing at 30%, and 10M doses of a vaccine per day are available after licensure, compared to the baseline case where no vaccine is ever approved.

	Vaccine Efficacy (%)			
	30		**50**	
	$\mathbb{E}[\Delta\text{Infections}]$	$\mathbb{E}[\Delta\text{Deaths}]$	$\mathbb{E}[\Delta\text{Infections}]$	$\mathbb{E}[\Delta\text{Deaths}]$
Status Quo				
RCT	437	4	3,735	30
ORCT	1,533	12	5,986	48
ARCT	592	4	5,288	42
HCT (30-day set-up)	3,419	28	142,814	1,141
HCT (60-day set-up)	2,637	21	111,554	891
HCT (90-day set-up)	2,037	17	87,572	699
HCT (120-day set-up)	1,572	13	69,039	551
Behavioral				
RCT	4,525	36	386,046	3,026
ORCT	30,524	243	1,102,052	8,763
ARCT	44,995	358	2,557,372	20,395
HCT (30-day set-up)	99,301	793	4,042,120	32,277
HCT (60-day set-up)	71,062	567	2,891,534	23,073
HCT (90-day set-up)	51,082	407	2,074,828	16,540
HCT (120-day set-up)	37,195	296	1,506,259	11,991
Ramp				
RCT	16,969	131	1,131,380	8,763
ORCT	88,322	700	2,719,614	21,513
ARCT	118,816	943	5,548,454	44,098
HCT (30-day set-up)	222,651	1,774	8,866,332	70,659
HCT (60-day set-up)	173,482	1,381	6,923,750	55,119
HCT (90-day set-up)	134,041	1,065	5,357,518	42,589
HCT (120-day set-up)	103,112	818	4,123,460	32,716

Table C.4 (continued): Expected number of incremental infections and deaths avoided in the U.S. under different trial designs, vaccine efficacies, and epidemiological scenarios, assuming trials start on August 1, 2020, superiority-by-margin testing at 30%, and 10M doses of a vaccine per day are available after licensure, compared to the baseline case in which no vaccine is ever approved.

	Vaccine Efficacy (%)			
	70		**90**	
	$\mathbb{E}[\Delta\text{Infections}]$	$\mathbb{E}[\Delta\text{Deaths}]$	$\mathbb{E}[\Delta\text{Infections}]$	$\mathbb{E}[\Delta\text{Deaths}]$
Status Quo				
RCT	13,837	109	21,254	168
ORCT	21,526	172	45,436	362
ARCT	28,115	223	65,559	522
HCT (30-day set-up)	156,885	1,254	159,876	1,277
HCT (60-day set-up)	122,482	979	124,777	997
HCT (90-day set-up)	96,111	768	97,886	782
HCT (120-day set-up)	75,747	605	77,132	615
Behavioral				
RCT	397,396	3,117	404,562	3,174
ORCT	1,425,995	11,345	1,457,500	11,598
ARCT	3,384,449	27,012	3,473,035	27,720
HCT (30-day set-up)	4,481,448	35,789	4,591,750	36,671
HCT (60-day set-up)	3,205,159	25,579	3,283,975	26,209
HCT (90-day set-up)	2,297,350	18,316	2,352,436	18,757
HCT (120-day set-up)	1,664,613	13,255	1,702,601	13,558
Ramp				
RCT	1,160,564	8,996	1,179,234	9,145
ORCT	3,969,592	31,468	4,050,013	32,111
ARCT	7,735,702	61,596	8,071,866	64,285
HCT (30-day set-up)	9,725,022	77,511	9,897,591	78,892
HCT (60-day set-up)	7,602,878	60,534	7,743,514	61,659
HCT (90-day set-up)	5,887,421	46,811	5,999,381	47,706
HCT (120-day set-up)	4,532,400	35,970	4,619,521	36,667

Table C.5: Expected number of incremental infections and deaths avoided in the U.S. under different trial designs, vaccine efficacies, and epidemiological scenarios, assuming trials start on August 1, 2020, superiority-by-margin testing at 30%, and infinite doses of a vaccine per day are available after licensure, compared to the baseline case in which no vaccine is ever approved.

	Vaccine Efficacy (%)			
	30		**50**	
	$\mathbb{E}[\Delta\text{Infections}]$	$\mathbb{E}[\Delta\text{Deaths}]$	$\mathbb{E}[\Delta\text{Infections}]$	$\mathbb{E}[\Delta\text{Deaths}]$
Status Quo				
RCT	489	4	4,111	33
ORCT	1,702	14	6,582	53
ARCT	662	5	5,825	46
HCT (30-day set-up)	3,835	31	158,149	1,263
HCT (60-day set-up)	2,956	24	123,272	984
HCT (90-day set-up)	2,282	19	96,592	771
HCT (120-day set-up)	1,762	14	76,033	607
Behavioral				
RCT	5,145	41	422,606	3,318
ORCT	34,444	275	1,231,877	9,802
ARCT	51,350	409	2,894,121	23,089
HCT (30-day set-up)	113,642	908	4,582,014	36,597
HCT (60-day set-up)	81,282	649	3,278,121	26,165
HCT (90-day set-up)	58,217	465	2,345,452	18,705
HCT (120-day set-up)	42,116	336	1,693,080	13,486
Ramp				
RCT	18,656	145	1,229,320	9,546
ORCT	98,315	780	3,001,533	23,767
ARCT	132,140	1,049	6,119,602	48,667
HCT (30-day set-up)	246,217	1,963	9,723,523	77,517
HCT (60-day set-up)	192,575	1,534	7,622,202	60,706
HCT (90-day set-up)	149,158	1,186	5,913,541	47,037
HCT (120-day set-up)	114,816	912	4,556,087	36,178

Table C.5 (continued): Expected number of incremental infections and deaths avoided in the U.S. under different trial designs, vaccine efficacies, and epidemiological scenarios, assuming trials start on August 1, 2020, superiority-by-margin testing at 30%, and infinite doses of a vaccine per day are available after licensure, compared to the baseline case in which no vaccine is ever approved.

	Vaccine Efficacy (%)			
	70		**90**	
	$\mathbb{E}[\Delta\text{Infections}]$	$\mathbb{E}[\Delta\text{Deaths}]$	$\mathbb{E}[\Delta\text{Infections}]$	$\mathbb{E}[\Delta\text{Deaths}]$
Status Quo				
RCT	15,118	120	23,096	183
ORCT	23,509	187	49,390	394
ARCT	30,759	245	71,348	568
HCT (30-day set-up)	172,598	1,380	174,917	1,398
HCT (60-day set-up)	134,478	1,075	136,254	1,088
HCT (90-day set-up)	105,338	841	106,709	852
HCT (120-day set-up)	82,896	662	83,965	670
Behavioral				
RCT	432,235	3,396	437,725	3,439
ORCT	1,587,079	12,634	1,613,158	12,843
ARCT	3,808,128	30,401	3,881,898	30,991
HCT (30-day set-up)	5,039,465	40,253	5,128,348	40,964
HCT (60-day set-up)	3,605,985	28,786	3,670,305	29,300
HCT (90-day set-up)	2,578,527	20,566	2,623,871	20,928
HCT (120-day set-up)	1,858,914	14,809	1,890,330	15,060
Ramp				
RCT	1,255,157	9,752	1,270,085	9,871
ORCT	4,358,075	34,576	4,422,914	35,094
ARCT	8,458,206	67,376	8,776,472	69,922
HCT (30-day set-up)	10,597,019	84,487	10,728,517	85,539
HCT (60-day set-up)	8,315,537	66,236	8,423,946	67,103
HCT (90-day set-up)	6,456,348	51,362	6,543,545	52,059
HCT (120-day set-up)	4,976,272	39,521	5,044,819	40,070

Table C.6: Expected number of incremental infections and deaths avoided in the U.S. under different trial designs, vaccine efficacies, and epidemiological scenarios, assuming trials start on August 1, 2020, superiority-by-margin testing at 50%, and 1M doses of a vaccine per day are available after licensure, compared to the baseline case where no vaccine is ever approved. We observe negative expected net values when vaccine efficacy is 30% because the candidate is almost never approved under superiority-by-margin testing. While a cost from conducting the trial is always incurred, the expected post-trial benefit is close to zero.

	Vaccine Efficacy (%)			
	30		**50**	
	$\mathbb{E}[\Delta\text{Infections}]$	$\mathbb{E}[\Delta\text{Deaths}]$	$\mathbb{E}[\Delta\text{Infections}]$	$\mathbb{E}[\Delta\text{Deaths}]$
Status Quo				
RCT	-34	0	319	3
ORCT	239	2	1,149	9
ARCT	-39	0	199	1
HCT (30-day set-up)	-171	-1	2,523	20
HCT (60-day set-up)	-171	-1	1,955	16
HCT (90-day set-up)	-171	-1	1,515	12
HCT (120-day set-up)	-171	-1	1,171	9
Behavioral				
RCT	-1,461	-11	2,242	17
ORCT	-331	-2	21,526	171
ARCT	-1,384	-11	29,583	235
HCT (30-day set-up)	-171	-1	67,258	537
HCT (60-day set-up)	-171	-1	48,652	388
HCT (90-day set-up)	-171	-1	35,595	283
HCT (120-day set-up)	-171	-1	26,494	210
Ramp				
RCT	-1,406	-11	10,693	82
ORCT	-198	-1	64,285	508
ARCT	-1,196	-9	82,127	649
HCT (30-day set-up)	-171	-1	164,007	1,305
HCT (60-day set-up)	-171	-1	127,036	1,009
HCT (90-day set-up)	-171	-1	98,023	777
HCT (120-day set-up)	-171	-1	75,645	598

Table C.6 (continued): Expected number of incremental infections and deaths avoided in the U.S. under different trial designs, vaccine efficacies, and epidemiological scenarios, assuming trials start on August 1, 2020, superiority-by-margin testing at 50%, and 1M doses of a vaccine per day are available after licensure, compared to the baseline case where no vaccine is ever approved.

	Vaccine Efficacy (%)			
	70		**90**	
	$\mathbb{E}[\Delta\text{Infections}]$	$\mathbb{E}[\Delta\text{Deaths}]$	$\mathbb{E}[\Delta\text{Infections}]$	$\mathbb{E}[\Delta\text{Deaths}]$
Status Quo				
RCT	4,091	32	14,935	118
ORCT	6,123	49	26,189	208
ARCT	3,840	30	27,107	215
HCT (30-day set-up)	113,800	910	120,945	966
HCT (60-day set-up)	89,345	713	94,885	758
HCT (90-day set-up)	70,439	562	74,766	597
HCT (120-day set-up)	55,722	445	59,123	471
Behavioral				
RCT	289,168	2,251	306,050	2,386
ORCT	955,088	7,581	1,065,183	8,459
ARCT	2,043,288	16,282	2,423,068	19,321
HCT (30-day set-up)	2,954,925	23,577	3,189,157	25,451
HCT (60-day set-up)	2,130,938	16,986	2,294,765	18,295
HCT (90-day set-up)	1,552,596	12,359	1,666,446	13,269
HCT (120-day set-up)	1,150,715	9,144	1,230,321	9,780
Ramp				
RCT	899,765	6,909	937,666	7,212
ORCT	2,467,656	19,477	3,047,293	24,089
ARCT	4,714,327	37,425	6,088,218	48,416
HCT (30-day set-up)	7,066,008	56,242	7,523,068	59,896
HCT (60-day set-up)	5,478,287	43,541	5,837,268	46,409
HCT (90-day set-up)	4,226,532	33,526	4,503,392	35,738
HCT (120-day set-up)	3,258,390	25,781	3,469,234	27,465

Table C.7: Expected number of incremental infections and deaths avoided in the U.S. under different trial designs, vaccine efficacies, and epidemiological scenarios, assuming trials start on August 1, 2020, superiority-by-margin testing at 50%, and 10M doses of a vaccine per day are available after licensure, compared to the baseline case where no vaccine is ever approved. We observe negative expected net values when vaccine efficacy is 30% because the candidate is almost never approved under superiority-by-margin testing. While a cost from conducting the trial is always incurred, the expected post-trial benefit is close to zero.

	Vaccine Efficacy (%)			
	30		**50**	
	$\mathbb{E}[\Delta\text{Infections}]$	$\mathbb{E}[\Delta\text{Deaths}]$	$\mathbb{E}[\Delta\text{Infections}]$	$\mathbb{E}[\Delta\text{Deaths}]$
Status Quo				
RCT	-25	0	471	4
ORCT	374	3	1,625	13
ARCT	-33	0	298	2
HCT (30-day set-up)	-171	-1	3,675	29
HCT (60-day set-up)	-171	-1	2,842	23
HCT (90-day set-up)	-171	-1	2,203	18
HCT (120-day set-up)	-171	-1	1,709	14
Behavioral				
RCT	-1,461	-11	3,852	30
ORCT	-331	-2	32,156	256
ARCT	-1,384	-11	46,267	368
HCT (30-day set-up)	-171	-1	107,601	859
HCT (60-day set-up)	-171	-1	76,935	614
HCT (90-day set-up)	-171	-1	55,168	440
HCT (120-day set-up)	-171	-1	40,014	319
Ramp				
RCT	-1,406	-11	14,720	115
ORCT	-183	-1	93,009	738
ARCT	-1,142	-9	119,304	947
HCT (30-day set-up)	-171	-1	236,179	1,882
HCT (60-day set-up)	-171	-1	184,404	1,468
HCT (90-day set-up)	-171	-1	142,660	1,134
HCT (120-day set-up)	-171	-1	109,769	871

Table C.7 (continued): Expected number of incremental infections and deaths avoided in the U.S. under different trial designs, vaccine efficacies, and epidemiological scenarios, assuming trials start on August 1, 2020, superiority-by-margin testing at 50%, and 10M doses of a vaccine per day are available after licensure, compared to the baseline case where no vaccine is ever approved.

	Vaccine Efficacy (%)			
	70		**90**	
	$\mathbb{E}[\Delta\text{Infections}]$	$\mathbb{E}[\Delta\text{Deaths}]$	$\mathbb{E}[\Delta\text{Infections}]$	$\mathbb{E}[\Delta\text{Deaths}]$
Status Quo				
RCT	5,536	44	19,507	154
ORCT	8,217	66	34,029	271
ARCT	5,170	41	35,268	280
HCT (30-day set-up)	155,455	1,243	159,876	1,277
HCT (60-day set-up)	121,365	970	124,777	997
HCT (90-day set-up)	95,234	761	97,886	782
HCT (120-day set-up)	75,056	599	77,132	615
Behavioral				
RCT	397,396	3,117	404,562	3,174
ORCT	1,352,103	10,757	1,457,500	11,598
ARCT	3,037,771	24,238	3,473,025	27,720
HCT (30-day set-up)	4,440,619	35,463	4,591,750	36,671
HCT (60-day set-up)	3,175,958	25,346	3,283,975	26,209
HCT (90-day set-up)	2,276,419	18,149	2,352,436	18,757
HCT (120-day set-up)	1,649,447	13,134	1,702,601	13,558
Ramp				
RCT	1,160,564	8,996	1,179,234	9,145
ORCT	3,387,704	26,840	4,050,013	32,111
ARCT	6,492,110	51,647	8,067,450	64,250
HCT (30-day set-up)	9,636,422	76,805	9,897,591	78,892
HCT (60-day set-up)	7,533,612	59,983	7,743,514	61,659
HCT (90-day set-up)	5,833,783	46,385	5,999,381	47,706
HCT (120-day set-up)	4,491,107	35,642	4,619,521	36,667

Table C.8: Expected number of incremental infections and deaths avoided in the U.S. under different trial designs, vaccine efficacies, and epidemiological scenarios, assuming trials start on August 1, 2020, superiority-by-margin testing at 50%, and infinite doses of a vaccine per day are available after licensure, compared to the baseline case where no vaccine is ever approved. We observe negative expected net values when vaccine efficacy is 30% because the candidate is almost never approved under superiority-by-margin testing. While a cost from conducting the trial is always incurred, the expected post-trial benefit is close to zero.

	Vaccine Efficacy (%)			
	30		**50**	
	$\mathbb{E}[\Delta\text{Infections}]$	$\mathbb{E}[\Delta\text{Deaths}]$	$\mathbb{E}[\Delta\text{Infections}]$	$\mathbb{E}[\Delta\text{Deaths}]$
Status Quo				
RCT	-22	0	523	4
ORCT	416	3	1,789	14
ARCT	-31	0	332	2
HCT (30-day set-up)	-171	-1	4,084	33
HCT (60-day set-up)	-171	-1	3,154	25
HCT (90-day set-up)	-171	-1	2,443	20
HCT (120-day set-up)	-171	-1	1,895	15
Behavioral				
RCT	-1,461	-11	4,337	34
ORCT	-331	-2	36,046	287
ARCT	-1,384	-11	52,340	417
HCT (30-day set-up)	-171	-1	121,991	974
HCT (60-day set-up)	-171	-1	87,239	696
HCT (90-day set-up)	-171	-1	62,381	498
HCT (120-day set-up)	-171	-1	44,993	358
Ramp				
RCT	-1,406	-11	16,101	126
ORCT	-178	-1	102,769	816
ARCT	-1,126	-9	131,636	1,045
HCT (30-day set-up)	-171	-1	259,025	2,065
HCT (60-day set-up)	-171	-1	203,020	1,617
HCT (90-day set-up)	-171	-1	157,479	1,253
HCT (120-day set-up)	-171	-1	121,300	963

Table C.8 (continued): Expected number of incremental infections and deaths avoided in the U.S. under different trial designs, vaccine efficacies, and epidemiological scenarios, assuming trials start on August 1, 2020, superiority-by-margin testing at 50%, and infinite doses of a vaccine per day are available after licensure, compared to the baseline case where no vaccine is ever approved.

	Vaccine Efficacy (%)			
	70		**90**	
	$\mathbb{E}[\Delta\text{Infections}]$	$\mathbb{E}[\Delta\text{Deaths}]$	$\mathbb{E}[\Delta\text{Infections}]$	$\mathbb{E}[\Delta\text{Deaths}]$
Status Quo				
RCT	6,050	48	21,198	168
ORCT	8,976	72	36,974	295
ARCT	5,655	45	38,342	305
HCT (30-day set-up)	171,025	1,367	174,917	1,398
HCT (60-day set-up)	133,252	1,065	136,254	1,088
HCT (90-day set-up)	104,377	833	106,709	852
HCT (120-day set-up)	82,140	656	83,965	670
Behavioral				
RCT	432,235	3,396	437,725	3,439
ORCT	1,504,842	11,979	1,613,158	12,843
ARCT	3,416,029	27,264	3,881,886	30,991
HCT (30-day set-up)	4,993,552	39,886	5,128,348	40,964
HCT (60-day set-up)	3,573,132	28,524	3,670,305	29,300
HCT (90-day set-up)	2,555,035	20,379	2,623,871	20,928
HCT (120-day set-up)	1,841,978	14,674	1,890,330	15,060
Ramp				
RCT	1,255,157	9,752	1,270,085	9,871
ORCT	3,718,588	29,487	4,422,914	35,094
ARCT	7,109,717	56,588	8,771,717	69,884
HCT (30-day set-up)	10,500,475	83,717	10,728,517	85,539
HCT (60-day set-up)	8,239,778	65,633	8,423,946	67,103
HCT (90-day set-up)	6,397,527	50,894	6,543,545	52,059
HCT (120-day set-up)	4,930,935	39,161	5,044,819	40,070

Table C.9: Estimated date of licensure and probability of approval under different trial designs, vaccine efficacies, and epidemiological scenarios, assuming trials start on August 1, 2020, superiority testing, and 1M doses of a vaccine per day are available after licensure. For ARCT, we report the median date of licensure over all Monte Carlo simulations. *Abbreviations*: DoL, date of licensure (month/day/year); PoA, probability of approval.

	Vaccine Efficacy (%)			
	30		**50**	
	DoL	PoA (%)	DoL	PoA (%)
Status Quo				
RCT	11/19/21	20.2	11/19/21	55.9
ORCT	08/14/21	13.6	08/15/21	38.9
ARCT	07/02/21	14.5	06/02/21	44.2
HCT (30-day set-up)	03/09/21	98.1	03/09/21	100.0
HCT (60-day set-up)	04/08/21	98.1	04/08/21	100.0
HCT (90-day set-up)	05/08/21	98.1	05/08/21	100.0
HCT (120-day set-up)	06/07/21	98.1	06/07/21	100.0
Behavioral				
RCT	11/19/21	100.0	11/19/21	100.0
ORCT	06/24/21	90.5	06/22/21	100.0
ARCT	04/03/21	100.0	04/03/21	100.0
HCT (30-day set-up)	03/09/21	98.1	03/09/21	100.0
HCT (60-day set-up)	04/08/21	98.1	04/08/21	100.0
HCT (90-day set-up)	05/08/21	98.1	05/08/21	100.0
HCT (120-day set-up)	06/07/21	98.1	06/07/21	100.0
Ramp				
RCT	11/19/21	100.0	11/19/21	100.0
ORCT	07/06/21	88.9	06/22/21	99.6
ARCT	05/03/21	100.0	04/03/21	100.0
HCT (30-day set-up)	03/09/21	98.1	03/09/21	100.0
HCT (60-day set-up)	04/08/21	98.1	04/08/21	100.0
HCT (90-day set-up)	05/08/21	98.1	05/08/21	100.0
HCT (120-day set-up)	06/07/21	98.1	06/07/21	100.0

Table C.9 (continued): Estimated date of licensure and probability of approval under different trial designs, vaccine efficacies, and epidemiological scenarios, assuming trials start on August 1, 2020, superiority testing, and 1M doses of a vaccine per day are available after licensure. *Abbreviations*: DoL, date of licensure (month/day/year); PoA, probability of approval.

	Vaccine Efficacy (%)			
	70		**90**	
	DoL	PoA (%)	DoL	PoA (%)
Status Quo				
RCT	11/19/21	89.9	11/19/21	99.6
ORCT	07/30/21	67.2	07/10/21	84.3
ARCT	06/02/21	83.8	06/02/21	99.6
HCT (30-day set-up)	03/09/21	100.0	03/09/21	100.0
HCT (60-day set-up)	04/08/21	100.0	04/08/21	100.0
HCT (90-day set-up)	05/08/21	100.0	05/08/21	100.0
HCT (120-day set-up)	06/07/21	100.0	06/07/21	100.0
Behavioral				
RCT	11/19/21	100.0	11/19/21	100.0
ORCT	06/22/21	100.0	06/22/21	100.0
ARCT	04/03/21	100.0	04/03/21	100.0
HCT (30-day set-up)	03/09/21	100.0	03/09/21	100.0
HCT (60-day set-up)	04/08/21	100.0	04/08/21	100.0
HCT (90-day set-up)	05/08/21	100.0	05/08/21	100.0
HCT (120-day set-up)	06/07/21	100.0	06/07/21	100.0
Ramp				
RCT	11/19/21	100.0	11/19/21	100.0
ORCT	06/22/21	100.0	06/22/21	100.0
ARCT	04/03/21	100.0	04/03/21	100.0
HCT (30-day set-up)	03/09/21	100.0	03/09/21	100.0
HCT (60-day set-up)	04/08/21	100.0	04/08/21	100.0
HCT (90-day set-up)	05/08/21	100.0	05/08/21	100.0
HCT (120-day set-up)	06/07/21	100.0	06/07/21	100.0

Table C.10: Estimated date of licensure and probability of approval under different trial designs, vaccine efficacies, and epidemiological scenarios, assuming trials start on August 1, 2020, superiority testing, and 10M doses of a vaccine per day are available after licensure. For ARCT, we report the median date of licensure over all Monte Carlo simulations. *Abbreviations*: DoL, date of licensure (month/day/year); PoA, probability of approval.

	Vaccine Efficacy (%)			
	30		**50**	
	DoL	PoA (%)	DoL	PoA (%)
Status Quo				
RCT	11/19/21	20.2	11/19/21	55.9
ORCT	08/15/21	13.8	08/15/21	38.9
ARCT	07/02/21	14.5	06/02/21	44.2
HCT (30-day set-up)	03/09/21	98.1	03/09/21	100.0
HCT (60-day set-up)	04/08/21	98.1	04/08/21	100.0
HCT (90-day set-up)	05/08/21	98.1	05/08/21	100.0
HCT (120-day set-up)	06/07/21	98.1	06/07/21	100.0
Behavioral				
RCT	11/19/21	100.0	11/19/21	100.0
ORCT	06/23/21	89.6	06/22/21	100.0
ARCT	04/03/21	100.0	04/03/21	100.0
HCT (30-day set-up)	03/09/21	98.1	03/09/21	100.0
HCT (60-day set-up)	04/08/21	98.1	04/08/21	100.0
HCT (90-day set-up)	05/08/21	98.1	05/08/21	100.0
HCT (120-day set-up)	06/07/21	98.1	06/07/21	100.0
Ramp				
RCT	11/19/21	100.0	11/19/21	100.0
ORCT	07/06/21	88.9	06/22/21	99.6
ARCT	05/03/21	100.0	04/03/21	100.0
HCT (30-day set-up)	03/09/21	98.1	03/09/21	100.0
HCT (60-day set-up)	04/08/21	98.1	04/08/21	100.0
HCT (90-day set-up)	05/08/21	98.1	05/08/21	100.0
HCT (120-day set-up)	06/07/21	98.1	06/07/21	100.0

Table C.10 (continued): Estimated date of licensure and probability of approval under different trial designs, vaccine efficacies, and epidemiological scenarios, assuming trials start on August 1, 2020, superiority testing, and 10M doses of a vaccine per day are available after licensure. *Abbreviations*: DoL, date of licensure (month/day/year); PoA, probability of approval.

	Vaccine Efficacy (%)			
	70		**90**	
	DoL	PoA (%)	DoL	PoA (%)
Status Quo				
RCT	11/19/21	89.9	11/19/21	99.6
ORCT	07/30/21	67.2	07/10/21	84.3
ARCT	06/02/21	83.8	06/02/21	99.6
HCT (30-day set-up)	03/09/21	100.0	03/09/21	100.0
HCT (60-day set-up)	04/08/21	100.0	04/08/21	100.0
HCT (90-day set-up)	05/08/21	100.0	05/08/21	100.0
HCT (120-day set-up)	06/07/21	100.0	06/07/21	100.0
Behavioral				
RCT	11/19/21	100.0	11/19/21	100.0
ORCT	06/22/21	100.0	06/22/21	100.0
ARCT	04/03/21	100.0	04/03/21	100.0
HCT (30-day set-up)	03/09/21	100.0	03/09/21	100.0
HCT (60-day set-up)	04/08/21	100.0	04/08/21	100.0
HCT (90-day set-up)	05/08/21	100.0	05/08/21	100.0
HCT (120-day set-up)	06/07/21	100.0	06/07/21	100.0
Ramp				
RCT	11/19/21	100.0	11/19/21	100.0
ORCT	06/22/21	100.0	06/22/21	100.0
ARCT	04/03/21	100.0	04/03/21	100.0
HCT (30-day set-up)	03/09/21	100.0	03/09/21	100.0
HCT (60-day set-up)	04/08/21	100.0	04/08/21	100.0
HCT (90-day set-up)	05/08/21	100.0	05/08/21	100.0
HCT (120-day set-up)	06/07/21	100.0	06/07/21	100.0

Table C.11: Estimated date of licensure and probability of approval under different trial designs, vaccine efficacies, and epidemiological scenarios, assuming trials start on August 1, 2020, superiority testing, and infinite doses of a vaccine per day are available after licensure. For ARCT, we report the median date of licensure over all Monte Carlo simulations. *Abbreviations*: DoL, date of licensure (month/day/year); PoA, probability of approval.

	Vaccine Efficacy (%)			
	30		**50**	
	DoL	PoA (%)	DoL	PoA (%)
Status Quo				
RCT	11/19/21	20.2	11/19/21	55.9
ORCT	08/14/21	13.6	08/14/21	38.6
ARCT	07/02/21	14.5	06/02/21	44.2
HCT (30-day set-up)	03/09/21	98.1	03/09/21	100.0
HCT (60-day set-up)	04/08/21	98.1	04/08/21	100.0
HCT (90-day set-up)	05/08/21	98.1	05/08/21	100.0
HCT (120-day set-up)	06/07/21	98.1	06/07/21	100.0
Behavioral				
RCT	11/19/21	100.0	11/19/21	100.0
ORCT	06/23/21	89.6	06/22/21	100.0
ARCT	04/03/21	100.0	04/03/21	100.0
HCT (30-day set-up)	03/09/21	98.1	03/09/21	100.0
HCT (60-day set-up)	04/08/21	98.1	04/08/21	100.0
HCT (90-day set-up)	05/08/21	98.1	05/08/21	100.0
HCT (120-day set-up)	06/07/21	98.1	06/07/21	100.0
Ramp				
RCT	11/19/21	100.0	11/19/21	100.0
ORCT	07/06/21	88.9	06/22/21	99.6
ARCT	05/03/21	100.0	04/03/21	100.0
HCT (30-day set-up)	03/09/21	98.1	03/09/21	100.0
HCT (60-day set-up)	04/08/21	98.1	04/08/21	100.0
HCT (90-day set-up)	05/08/21	98.1	05/08/21	100.0
HCT (120-day set-up)	06/07/21	98.1	06/07/21	100.0

Table C.11 (continued): Estimated date of licensure and probability of approval under different trial designs, vaccine efficacies, and epidemiological scenarios, assuming trials start on August 1, 2020, superiority testing, and infinite doses of a vaccine per day are available after licensure. *Abbreviations*: DoL, date of licensure (month/day/year); PoA, probability of approval.

	Vaccine Efficacy (%)			
	70		**90**	
	DoL	PoA (%)	DoL	PoA (%)
Status Quo				
RCT	11/19/21	89.9	11/19/21	99.6
ORCT	07/30/21	67.2	07/10/21	84.3
ARCT	06/02/21	83.8	06/02/21	99.6
HCT (30-day set-up)	03/09/21	100.0	03/09/21	100.0
HCT (60-day set-up)	04/08/21	100.0	04/08/21	100.0
HCT (90-day set-up)	05/08/21	100.0	05/08/21	100.0
HCT (120-day set-up)	06/07/21	100.0	06/07/21	100.0
Behavioral				
RCT	11/19/21	100.0	11/19/21	100.0
ORCT	06/22/21	100.0	06/22/21	100.0
ARCT	04/03/21	100.0	04/03/21	100.0
HCT (30-day set-up)	03/09/21	100.0	03/09/21	100.0
HCT (60-day set-up)	04/08/21	100.0	04/08/21	100.0
HCT (90-day set-up)	05/08/21	100.0	05/08/21	100.0
HCT (120-day set-up)	06/07/21	100.0	06/07/21	100.0
Ramp				
RCT	11/19/21	100.0	11/19/21	100.0
ORCT	06/22/21	100.0	06/22/21	100.0
ARCT	04/03/21	100.0	04/03/21	100.0
HCT (30-day set-up)	03/09/21	100.0	03/09/21	100.0
HCT (60-day set-up)	04/08/21	100.0	04/08/21	100.0
HCT (90-day set-up)	05/08/21	100.0	05/08/21	100.0
HCT (120-day set-up)	06/07/21	100.0	06/07/21	100.0

Table C.12: Estimated date of licensure and probability of approval under different trial designs, vaccine efficacies, and epidemiological scenarios, assuming trials start on August 1, 2020, superiority-by-margin testing at 30%, and 1M doses of a vaccine per day are available after licensure. For ARCT, we report the median date of licensure over all Monte Carlo simulations. *Abbreviations*: DoL, date of licensure (month/day/year); PoA, probability of approval.

	Vaccine Efficacy (%)			
	30		**50**	
	DoL	PoA (%)	DoL	PoA (%)
Status Quo				
RCT	11/19/21	2.5	11/19/21	18.2
ORCT	06/22/21	2.5	06/22/21	13.6
ARCT	07/02/21	1.0	07/02/21	8.3
HCT (30-day set-up)	03/09/21	2.5	03/09/21	93.8
HCT (60-day set-up)	04/08/21	2.5	04/08/21	93.8
HCT (90-day set-up)	05/08/21	2.5	05/08/21	93.8
HCT (120-day set-up)	06/07/21	2.5	06/07/21	93.8
Behavioral				
RCT	11/19/21	1.6	11/19/21	100.0
ORCT	06/22/21	2.4	06/22/21	78.3
ARCT	07/22/21	2.4	05/03/21	100.0
HCT (30-day set-up)	03/09/21	2.5	03/09/21	93.8
HCT (60-day set-up)	04/08/21	2.5	04/08/21	93.8
HCT (90-day set-up)	05/08/21	2.5	05/08/21	93.8
HCT (120-day set-up)	06/07/21	2.5	06/07/21	93.8
Ramp				
RCT	11/19/21	1.7	11/19/21	100.0
ORCT	06/22/21	2.4	06/22/21	61.3
ARCT	08/21/21	2.6	05/03/21	99.9
HCT (30-day set-up)	03/09/21	2.5	03/09/21	93.8
HCT (60-day set-up)	04/08/21	2.5	04/08/21	93.8
HCT (90-day set-up)	05/08/21	2.5	05/08/21	93.8
HCT (120-day set-up)	06/07/21	2.5	06/07/21	93.8

Table C.12 (continued): Estimated date of licensure and probability of approval under different trial designs, vaccine efficacies, and epidemiological scenarios, assuming trials start on August 1, 2020, superiority-by-margin testing at 30%, and 1M doses of a vaccine per day are available after licensure. *Abbreviations*: DoL, date of licensure (month/day/year); PoA, probability of approval.

	Vaccine Efficacy (%)			
	70		**90**	
	DoL	PoA (%)	DoL	PoA (%)
Status Quo				
RCT	11/19/21	65.1	11/19/21	98.2
ORCT	08/06/21	42.7	07/31/21	75.9
ARCT	07/02/21	42.5	07/02/21	94.2
HCT (30-day set-up)	03/09/21	100.0	03/09/21	100.0
HCT (60-day set-up)	04/08/21	100.0	04/08/21	100.0
HCT (90-day set-up)	05/08/21	100.0	05/08/21	100.0
HCT (120-day set-up)	06/07/21	100.0	06/07/21	100.0
Behavioral				
RCT	11/19/21	100.0	11/19/21	100.0
ORCT	06/22/21	100.0	06/22/21	100.0
ARCT	04/03/21	100.0	04/03/21	100.0
HCT (30-day set-up)	03/09/21	100.0	03/09/21	100.0
HCT (60-day set-up)	04/08/21	100.0	04/08/21	100.0
HCT (90-day set-up)	05/08/21	100.0	05/08/21	100.0
HCT (120-day set-up)	06/07/21	100.0	06/07/21	100.0
Ramp				
RCT	11/19/21	100.0	11/19/21	100.0
ORCT	06/30/21	99.9	06/22/21	100.0
ARCT	04/03/21	100.0	04/03/21	100.0
HCT (30-day set-up)	03/09/21	100.0	03/09/21	100.0
HCT (60-day set-up)	04/08/21	100.0	04/08/21	100.0
HCT (90-day set-up)	05/08/21	100.0	05/08/21	100.0
HCT (120-day set-up)	06/07/21	100.0	06/07/21	100.0

Table C.13: Estimated date of licensure and probability of approval under different trial designs, vaccine efficacies, and epidemiological scenarios, assuming trials start on August 1, 2020, superiority-by-margin testing at 30%, and 10M doses of a vaccine per day are available after licensure. For ARCT, we report the median date of licensure over all Monte Carlo simulations. *Abbreviations*: DoL, date of licensure (month/day/year); PoA, probability of approval.

	Vaccine Efficacy (%)			
	30		**50**	
	DoL	PoA (%)	DoL	PoA (%)
Status Quo				
RCT	11/19/21	2.5	11/19/21	18.2
ORCT	06/22/21	2.5	06/22/21	13.6
ARCT	07/02/21	1.0	07/02/21	8.3
HCT (30-day set-up)	03/09/21	2.5	03/09/21	93.8
HCT (60-day set-up)	04/08/21	2.5	04/08/21	93.8
HCT (90-day set-up)	05/08/21	2.5	05/08/21	93.8
HCT (120-day set-up)	06/07/21	2.5	06/07/21	93.8
Behavioral				
RCT	11/19/21	1.6	11/19/21	100.0
ORCT	06/22/21	2.4	06/22/21	78.3
ARCT	07/22/21	2.4	05/03/21	100.0
HCT (30-day set-up)	03/09/21	2.5	03/09/21	93.8
HCT (60-day set-up)	04/08/21	2.5	04/08/21	93.8
HCT (90-day set-up)	05/08/21	2.5	05/08/21	93.8
HCT (120-day set-up)	06/07/21	2.5	06/07/21	93.8
Ramp				
RCT	11/19/21	1.7	11/19/21	100.0
ORCT	06/22/21	2.4	06/22/21	61.3
ARCT	08/21/21	2.6	05/03/21	99.9
HCT (30-day set-up)	03/09/21	2.5	03/09/21	93.8
HCT (60-day set-up)	04/08/21	2.5	04/08/21	93.8
HCT (90-day set-up)	05/08/21	2.5	05/08/21	93.8
HCT (120-day set-up)	06/07/21	2.5	06/07/21	93.8

Table C.13 (continued): Estimated date of licensure and probability of approval under different trial designs, vaccine efficacies, and epidemiological scenarios, assuming trials start on August 1, 2020, superiority-by-margin testing at 30%, and 10M doses of a vaccine per day are available after licensure. *Abbreviations*: DoL, date of licensure (month/day/year); PoA, probability of approval.

	Vaccine Efficacy (%)			
	70		**90**	
	DoL	PoA (%)	DoL	PoA (%)
Status Quo				
RCT	11/19/21	65.1	11/19/21	98.2
ORCT	08/06/21	42.7	07/31/21	75.9
ARCT	07/02/21	42.5	07/02/21	94.2
HCT (30-day set-up)	03/09/21	100.0	03/09/21	100.0
HCT (60-day set-up)	04/08/21	100.0	04/08/21	100.0
HCT (90-day set-up)	05/08/21	100.0	05/08/21	100.0
HCT (120-day set-up)	06/07/21	100.0	06/07/21	100.0
Behavioral				
RCT	11/19/21	100.0	11/19/21	100.0
ORCT	06/22/21	100.0	06/22/21	100.0
ARCT	04/03/21	100.0	04/03/21	100.0
HCT (30-day set-up)	03/09/21	100.0	03/09/21	100.0
HCT (60-day set-up)	04/08/21	100.0	04/08/21	100.0
HCT (90-day set-up)	05/08/21	100.0	05/08/21	100.0
HCT (120-day set-up)	06/07/21	100.0	06/07/21	100.0
Ramp				
RCT	11/19/21	100.0	11/19/21	100.0
ORCT	06/29/21	99.9	06/22/21	100.0
ARCT	04/03/21	100.0	04/03/21	100.0
HCT (30-day set-up)	03/09/21	100.0	03/09/21	100.0
HCT (60-day set-up)	04/08/21	100.0	04/08/21	100.0
HCT (90-day set-up)	05/08/21	100.0	05/08/21	100.0
HCT (120-day set-up)	06/07/21	100.0	06/07/21	100.0

Table C.14: Estimated date of licensure and probability of approval under different trial designs, vaccine efficacies, and epidemiological scenarios, assuming trials start on August 1, 2020, superiority-by-margin testing at 30%, and infinite doses of a vaccine per day are available after licensure. For ARCT, we report the median date of licensure over all Monte Carlo simulations. *Abbreviations*: DoL, date of licensure (month/day/year); PoA, probability of approval.

	Vaccine Efficacy (%)			
	30		**50**	
	DoL	PoA (%)	DoL	PoA (%)
Status Quo				
RCT	11/19/21	2.5	11/19/21	18.2
ORCT	06/22/21	2.5	06/22/21	13.5
ARCT	07/02/21	1.0	07/02/21	8.3
HCT (30-day set-up)	03/09/21	2.5	03/09/21	93.8
HCT (60-day set-up)	04/08/21	2.5	04/08/21	93.8
HCT (90-day set-up)	05/08/21	2.5	05/08/21	93.8
HCT (120-day set-up)	06/07/21	2.5	06/07/21	93.8
Behavioral				
RCT	11/19/21	1.6	11/19/21	100.0
ORCT	06/22/21	2.4	06/22/21	78.3
ARCT	07/22/21	2.4	05/03/21	100.0
HCT (30-day set-up)	03/09/21	2.5	03/09/21	93.8
HCT (60-day set-up)	04/08/21	2.5	04/08/21	93.8
HCT (90-day set-up)	05/08/21	2.5	05/08/21	93.8
HCT (120-day set-up)	06/07/21	2.5	06/07/21	93.8
Ramp				
RCT	11/19/21	1.7	11/19/21	100.0
ORCT	06/22/21	2.4	06/22/21	61.3
ARCT	08/21/21	2.6	05/03/21	99.9
HCT (30-day set-up)	03/09/21	2.5	03/09/21	93.8
HCT (60-day set-up)	04/08/21	2.5	04/08/21	93.8
HCT (90-day set-up)	05/08/21	2.5	05/08/21	93.8
HCT (120-day set-up)	06/07/21	2.5	06/07/21	93.8

Table C.14 (continued): Estimated date of licensure and probability of approval under different trial designs, vaccine efficacies, and epidemiological scenarios, assuming trials start on August 1, 2020, superiority-by-margin testing at 30%, and infinite doses of a vaccine per day are available after licensure. *Abbreviations*: DoL, date of licensure (month/day/year); PoA, probability of approval.

	Vaccine Efficacy (%)			
	70		**90**	
	DoL	PoA (%)	DoL	PoA (%)
Status Quo				
RCT	11/19/21	65.1	11/19/21	98.2
ORCT	08/06/21	42.7	07/31/21	75.9
ARCT	07/02/21	42.5	07/02/21	94.2
HCT (30-day set-up)	03/09/21	100.0	03/09/21	100.0
HCT (60-day set-up)	04/08/21	100.0	04/08/21	100.0
HCT (90-day set-up)	05/08/21	100.0	05/08/21	100.0
HCT (120-day set-up)	06/07/21	100.0	06/07/21	100.0
Behavioral				
RCT	11/19/21	100.0	11/19/21	100.0
ORCT	06/22/21	100.0	06/22/21	100.0
ARCT	04/03/21	100.0	04/03/21	100.0
HCT (30-day set-up)	03/09/21	100.0	03/09/21	100.0
HCT (60-day set-up)	04/08/21	100.0	04/08/21	100.0
HCT (90-day set-up)	05/08/21	100.0	05/08/21	100.0
HCT (120-day set-up)	06/07/21	100.0	06/07/21	100.0
Ramp				
RCT	11/19/21	100.0	11/19/21	100.0
ORCT	06/29/21	99.9	06/22/21	100.0
ARCT	04/03/21	100.0	04/03/21	100.0
HCT (30-day set-up)	03/09/21	100.0	03/09/21	100.0
HCT (60-day set-up)	04/08/21	100.0	04/08/21	100.0
HCT (90-day set-up)	05/08/21	100.0	05/08/21	100.0
HCT (120-day set-up)	06/07/21	100.0	06/07/21	100.0

Table C.15: Estimated date of licensure and probability of approval under different trial designs, vaccine efficacies, and epidemiological scenarios, assuming trials start on August 1, 2020, superiority-by-margin testing at 50%, and 1M doses of a vaccine per day are available after licensure. For ARCT, we report the median date of licensure over all Monte Carlo simulations. A blank entry indicates that the vaccine candidate is never approved. *Abbreviations*: DoL, date of licensure (month/day/year); PoA, probability of approval.

	Vaccine Efficacy (%)			
	30		**50**	
	DoL	PoA (%)	DoL	PoA (%)
Status Quo				
RCT	11/19/21	0.1	11/19/21	2.5
ORCT	06/22/21	0.3	06/22/21	2.5
ARCT		0.0	07/02/21	0.6
HCT (30-day set-up)		0.0	03/09/21	2.5
HCT (60-day set-up)		0.0	04/08/21	2.5
HCT (90-day set-up)		0.0	05/08/21	2.5
HCT (120-day set-up)		0.0	06/07/21	2.5
Behavioral				
RCT		0.0	11/19/21	1.3
ORCT		0.0	06/22/21	2.4
ARCT		0.0	06/02/21	2.4
HCT (30-day set-up)		0.0	03/09/21	2.5
HCT (60-day set-up)		0.0	04/08/21	2.5
HCT (90-day set-up)		0.0	05/08/21	2.5
HCT (120-day set-up)		0.0	06/07/21	2.5
Ramp				
RCT		0.0	11/19/21	1.4
ORCT		0.0	06/22/21	2.4
ARCT		0.0	06/02/21	2.5
HCT (30-day set-up)		0.0	03/09/21	2.5
HCT (60-day set-up)		0.0	04/08/21	2.5
HCT (90-day set-up)		0.0	05/08/21	2.5
HCT (120-day set-up)		0.0	06/07/21	2.5

Table C.15 (continued): Estimated date of licensure and probability of approval under different trial designs, vaccine efficacies, and epidemiological scenarios, assuming trials start on August 1, 2020, superiority-by-margin testing at 50%, and 1M doses of a vaccine per day are available after licensure. *Abbreviations*: DoL, date of licensure (month/day/year); PoA, probability of approval.

	Vaccine Efficacy (%)			
	70		**90**	
	DoL	PoA (%)	DoL	PoA (%)
Status Quo				
RCT	11/19/21	26.2	11/19/21	90.1
ORCT	08/06/21	16.3	07/31/21	53.5
ARCT	08/01/21	9.3	08/01/21	64.3
HCT (30-day set-up)	03/09/21	99.1	03/09/21	100.0
HCT (60-day set-up)	04/08/21	99.1	04/08/21	100.0
HCT (90-day set-up)	05/08/21	99.1	05/08/21	100.0
HCT (120-day set-up)	06/07/21	99.1	06/07/21	100.0
Behavioral				
RCT	11/19/21	100.0	11/19/21	100.0
ORCT	06/22/21	94.8	06/22/21	100.0
ARCT	04/03/21	100.0	04/03/21	100.0
HCT (30-day set-up)	03/09/21	99.1	03/09/21	100.0
HCT (60-day set-up)	04/08/21	99.1	04/08/21	100.0
HCT (90-day set-up)	05/08/21	99.1	05/08/21	100.0
HCT (120-day set-up)	06/07/21	99.1	06/07/21	100.0
Ramp				
RCT	11/19/21	100.0	11/19/21	100.0
ORCT	06/30/21	83.2	06/22/21	100.0
ARCT	05/03/21	100.0	04/03/21	100.0
HCT (30-day set-up)	03/09/21	99.1	03/09/21	100.0
HCT (60-day set-up)	04/08/21	99.1	04/08/21	100.0
HCT (90-day set-up)	05/08/21	99.1	05/08/21	100.0
HCT (120-day set-up)	06/07/21	99.1	06/07/21	100.0

Table C.16: Estimated date of licensure and probability of approval under different trial designs, vaccine efficacies, and epidemiological scenarios, assuming trials start on August 1, 2020, superiority-by-margin testing at 50%, and 10M doses of a vaccine per day are available after licensure. For ARCT, we report the median date of licensure over all Monte Carlo simulations. A blank entry indicates that the vaccine candidate is never approved. *Abbreviations*: DoL, date of licensure (month/day/year); PoA, probability of approval.

	Vaccine Efficacy (%)			
	30		**50**	
	DoL	PoA (%)	DoL	PoA (%)
Status Quo				
RCT	11/19/21	0.1	11/19/21	2.5
ORCT	06/22/21	0.3	06/22/21	2.5
ARCT		0.0	07/02/21	0.6
HCT (30-day set-up)		0.0	03/09/21	2.5
HCT (60-day set-up)		0.0	04/08/21	2.5
HCT (90-day set-up)		0.0	05/08/21	2.5
HCT (120-day set-up)		0.0	06/07/21	2.5
Behavioral				
RCT		0.0	11/19/21	1.3
ORCT		0.0	06/22/21	2.4
ARCT		0.0	06/02/21	2.4
HCT (30-day set-up)		0.0	03/09/21	2.5
HCT (60-day set-up)		0.0	04/08/21	2.5
HCT (90-day set-up)		0.0	05/08/21	2.5
HCT (120-day set-up)		0.0	06/07/21	2.5
Ramp				
RCT		0.0	11/19/21	1.4
ORCT		0.0	06/22/21	2.4
ARCT		0.0	06/02/21	2.5
HCT (30-day set-up)		0.0	03/09/21	2.5
HCT (60-day set-up)		0.0	04/08/21	2.5
HCT (90-day set-up)		0.0	05/08/21	2.5
HCT (120-day set-up)		0.0	06/07/21	2.5

Table C.16 (continued): Estimated date of licensure and probability of approval under different trial designs, vaccine efficacies, and epidemiological scenarios, assuming trials start on August 1, 2020, superiority-by-margin testing at 50%, and 10M doses of a vaccine per day are available after licensure. *Abbreviations*: DoL, date of licensure (month/day/year); PoA, probability of approval.

	Vaccine Efficacy (%)			
	70		**90**	
	DoL	PoA (%)	DoL	PoA (%)
Status Quo				
RCT	11/19/21	26.2	11/19/21	90.1
ORCT	08/06/21	16.3	07/31/21	53.5
ARCT	08/01/21	9.3	08/01/21	64.3
HCT (30-day set-up)	03/09/21	99.1	03/09/21	100.0
HCT (60-day set-up)	04/08/21	99.1	04/08/21	100.0
HCT (90-day set-up)	05/08/21	99.1	05/08/21	100.0
HCT (120-day set-up)	06/07/21	99.1	06/07/21	100.0
Behavioral				
RCT	11/19/21	100.0	11/19/21	100.0
ORCT	06/22/21	94.8	06/22/21	100.0
ARCT	04/03/21	100.0	04/03/21	100.0
HCT (30-day set-up)	03/09/21	99.1	03/09/21	100.0
HCT (60-day set-up)	04/08/21	99.1	04/08/21	100.0
HCT (90-day set-up)	05/08/21	99.1	05/08/21	100.0
HCT (120-day set-up)	06/07/21	99.1	06/07/21	100.0
Ramp				
RCT	11/19/21	100.0	11/19/21	100.0
ORCT	06/29/21	83.2	06/22/21	100.0
ARCT	05/03/21	100.0	04/03/21	100.0
HCT (30-day set-up)	03/09/21	99.1	03/09/21	100.0
HCT (60-day set-up)	04/08/21	99.1	04/08/21	100.0
HCT (90-day set-up)	05/08/21	99.1	05/08/21	100.0
HCT (120-day set-up)	06/07/21	99.1	06/07/21	100.0

Table C.17: Estimated date of licensure and probability of approval under different trial designs, vaccine efficacies, and epidemiological scenarios, assuming trials start on August 1, 2020, superiority-by-margin testing at 50%, and infinite doses of a vaccine per day are available after licensure. For ARCT, we report the median date of licensure over all Monte Carlo simulations. A blank entry indicates that the vaccine candidate is never approved. *Abbreviations*: DoL, date of licensure (month/day/year); PoA, probability of approval.

	Vaccine Efficacy (%)			
	30		**50**	
	DoL	PoA (%)	DoL	PoA (%)
Status Quo				
RCT	11/19/21	0.1	11/19/21	2.5
ORCT	06/22/21	0.3	06/22/21	2.5
ARCT		0.0	07/02/21	0.6
HCT (30-day set-up)		0.0	03/09/21	2.5
HCT (60-day set-up)		0.0	04/08/21	2.5
HCT (90-day set-up)		0.0	05/08/21	2.5
HCT (120-day set-up)		0.0	06/07/21	2.5
Behavioral				
RCT		0.0	11/19/21	1.3
ORCT		0.0	06/22/21	2.4
ARCT		0.0	06/02/21	2.4
HCT (30-day set-up)		0.0	03/09/21	2.5
HCT (60-day set-up)		0.0	04/08/21	2.5
HCT (90-day set-up)		0.0	05/08/21	2.5
HCT (120-day set-up)		0.0	06/07/21	2.5
Ramp				
RCT		0.0	11/19/21	1.4
ORCT		0.0	06/22/21	2.4
ARCT		0.0	06/02/21	2.5
HCT (30-day set-up)		0.0	03/09/21	2.5
HCT (60-day set-up)		0.0	04/08/21	2.5
HCT (90-day set-up)		0.0	05/08/21	2.5
HCT (120-day set-up)		0.0	06/07/21	2.5

Table C.17 (continued): Estimated date of licensure and probability of approval under different trial designs, vaccine efficacies, and epidemiological scenarios, assuming trials start on August 1, 2020, superiority-by-margin testing at 50%, and infinite doses of a vaccine per day are available after licensure. *Abbreviations*: DoL, date of licensure (month/day/year); PoA, probability of approval.

	Vaccine Efficacy (%)			
	70		**90**	
	DoL	PoA (%)	DoL	PoA (%)
Status Quo				
RCT	11/19/21	26.2	11/19/21	90.1
ORCT	08/06/21	16.3	07/31/21	53.5
ARCT	08/01/21	9.3	08/01/21	64.3
HCT (30-day set-up)	03/09/21	99.1	03/09/21	100.0
HCT (60-day set-up)	04/08/21	99.1	04/08/21	100.0
HCT (90-day set-up)	05/08/21	99.1	05/08/21	100.0
HCT (120-day set-up)	06/07/21	99.1	06/07/21	100.0
Behavioral				
RCT	11/19/21	100.0	11/19/21	100.0
ORCT	06/22/21	94.8	06/22/21	100.0
ARCT	04/03/21	100.0	04/03/21	100.0
HCT (30-day set-up)	03/09/21	99.1	03/09/21	100.0
HCT (60-day set-up)	04/08/21	99.1	04/08/21	100.0
HCT (90-day set-up)	05/08/21	99.1	05/08/21	100.0
HCT (120-day set-up)	06/07/21	99.1	06/07/21	100.0
Ramp				
RCT	11/19/21	100.0	11/19/21	100.0
ORCT	06/29/21	83.2	06/22/21	100.0
ARCT	05/03/21	100.0	04/03/21	100.0
HCT (30-day set-up)	03/09/21	99.1	03/09/21	100.0
HCT (60-day set-up)	04/08/21	99.1	04/08/21	100.0
HCT (90-day set-up)	05/08/21	99.1	05/08/21	100.0
HCT (120-day set-up)	06/07/21	99.1	06/07/21	100.0

C.7 Steps in HCT Setup

Steps in HCT setup include:

- Selection of SARS-CoV-2 challenge strain (assuming a currently circulating and predominant wild-type strain) with careful validation of provenance and health status of the subject from which the strain is procured or generation of viral strain by reverse genetics
- Selection of a high-level containment laboratory to prepare and manufacture the challenge strain, and contracting with said laboratory
- Purification and full characterization of challenge strain
- cGMP (current Good Manufacturing Practice) production of challenge pool
- Testing of challenge pool for impurities (including contaminating organisms)
- Titration of challenge strain in cell
- Development of clinical study protocol (design, inclusion/exclusion criteria, study endpoints)
- Validation of virologic and immunologic assays to be used in the clinical study
- Development of informed consent form, and compensation to be paid to volunteers
- Development of robust rescue protocols (supportive care, therapeutics)
- Regulatory approvals of each stage of the above steps submitted to FDA in an IND for 1) the challenge pool and separately 2) for the clinical study protocol, for their review
- Adaptation/development of a secure quarantine facility in a hospital setting with monitoring equipment, ventilation controls, and specialist staff
- Training of nurses, securing PPE and other equipment

- IRB review and approval of protocol
- Development of communications program, including dedicated website for sign-ons
- Recruitment of volunteers
- Intensive screening of volunteers for susceptibility to SARS-CoV-2, including prior exposure to human coronaviruses, known risk factors, including comorbidities, preexisting conditions, known genetic risk factors for severe COVID-19, and anti-interferon antibodies
- Final go-ahead from study sponsor and regulatory authority
- Conduct dose-ranging study to determine the lowest infectious dose/appropriate inoculum to reliably infect susceptible volunteers with challenge virus before proceeding with vaccinating and challenging volunteers per updated/revised study protocol

Appendix D

Supplement to Chapter 5

D.1 Company Screening

Therapeutics companies were identified from an initial list of 225 life sciences companies that have licensed IP from MIT. (For purposes of exposition, we will refer to a firm that licensed MIT IP as an "MIT licensee.") Based on company business descriptions and financial filings, 73 MIT licensees were identified by the authors as therapeutics companies, and further characterized as private versus public, and as alive versus acquired versus bankrupt.

Among the MIT licensees in therapeutics, Cell Genesys launched two spinoffs, Abgenix in 1996 and Ceregene in 2001. Abgenix eventually underwent its own IPO process in 1998, while Ceregene stayed private. Scios Nova also launched a spinoff, Guilford Pharmaceuticals in 1993, which went public in 1994. These three companies were added to the dataset, bringing the total count to 76 companies.

D.2 Orange Book Citations

Following the approach used by Stevens et al. [155], we use two databases to link MIT IP and drugs approved by the FDA. Our primary source is the Orange Book that is published and updated periodically by the FDA [154]. The publication identifies drug products approved on the basis of safety and effectiveness by the FDA under the Federal Food, Drug, and Cosmetic Act. It includes only currently marketed prescription drug products approved through NDAs and Abbreviated New Drug Applications (ANDAs or Generics) [222]. In addition to therapeutic equivalence evaluations, the Orange Book lists all patents protecting each approved product, as provided by the drug application owner.

First, we compile historical snapshots of the Orange Book published between 1985–2017 (cutoff at December 2017), made available by Jean Roth and Heidi Williams on the National Bureau of Economic Research data archive [164, 165]. To identify drugs that owe their origin, at least in part, to MIT IP, we search the compiled Orange Book dataset for applications that cite patents assigned to MIT. We use the USPTO Patent Full-Text and Patent Assignment databases to determine the chain of ownership for each patent [166, 167].

We find six NDAs in the Orange Book that cited MIT patents (see Table D.1). They correspond to five small molecule drugs and two in vivo diagnostic products. We note that latter, Cardiolite and Miraluma, are radioactive tracers used in nuclear medicine imaging. They fall outside the scope of our analysis which focuses on therapeutic drug discovery. Alnylam received FDA approval for patisiran, a novel first-in-class siRNA therapeutic for hereditary ATTR amyloidosis, in August 2018 (after the cutoff of our analysis, December 2017) [179]. Therefore, we excluded these drugs from the main text. Stevens at el. [155] credited the approval of Visudyne, a photodynamic therapy, to MIT. However, we did not find any MIT patent citations in Visudyne's NDA in the Orange Book. Most of Visudyne's cited patents were assigned to the University of British Columbia, QLT Phototherapeutics, and the Massachusetts Eye and Ear Infirmary.

Table D.1: Expanded list of MIT IP citations in the Orange Book.

NDA	Drug	Type	Patent No.	Title
18936	Sarafem	Small molecule drug	4,035,511	Process for promoting analgesia
18936	Sarafem	Small molecule drug	4,083,982	Process for producing analgesia
18936	Sarafem	Small molecule drug	4,971,998	Methods for treating the premenstrual or late luteal phase syndrome
19785	Cardiolite	In vivo diagnostics	4,452,774	Isonitrile radionuclide complexes for labelling and imaging agents
19785	Miraluma	In vivo diagnostics	4,452,774	Isonitrile radionuclide complexes for labelling and imaging agents
20344	Redux	Small molecule drug	4,309,445	d-Fenfluramine for modifying feeding behavior
20637	Gliadel	Small molecule drug	4,757,128	High molecular weight polyanhydride and preparation thereof
20637	Gliadel	Small molecule drug	4,789,724	Preparation of anhydride copolymers
20637	Gliadel	Small molecule drug	5,179,189	Fatty acid terminated polyanhydrides
207958	Spritam	Small molecule drug	6,471,992	Dosage form exhibiting rapid disperse properties, methods of use and process for the manufacture of same
207958	Spritam	Small molecule drug	9,463,160	Dosage form exhibiting rapid disperse properties, methods of use and process for the manufacture of same
210922	Patisiran	Small molecule drug	8,362,231	RNA interference mediating small RNA molecules
210922	Patisiran	Small molecule drug	8,372,968	RNA interference mediating small RNA molecules
210922	Patisiran	Small molecule drug	8,552,171	RNA sequence-specific mediators of RNA interference
210922	Patisiran	Small molecule drug	8,778,902	RNA interference mediating small RNA molecules
210922	Patisiran	Small molecule drug	8,895,718	RNA interference mediating small RNA molecules
210922	Patisiran	Small molecule drug	8,895,721	RNA interference mediating small RNA molecules
210922	Patisiran	Small molecule drug	9,193,753	RNA sequence-specific mediators of RNA interference
210922	Patisiran	Small molecule drug	9,567,582	RNA interference mediating small RNA molecules

D.3 Initial Public Offerings

We plot the capital generated by the IPOs of MIT licensees and the resulting shareholder dilution over time (see Figs. D-1 to D-3). Over the studied period, MIT licensees raised more capital, though this also came with increased dilution to shareholders. While inflation has been a likely contributor over the last two decades, Fig. D-2 highlights that it was not the sole contributor. After adjusting for inflation using the BRDPI, net proceeds experienced a high level of growth, an indication of the dramatic expansion of the biotech capital markets over the past two decades. In particular, these markets experienced a resurgence in the years leading up to the 2016 crash. The IPOs occurred in clusters that coincided with this window and other periods of favorable market conditions, as shown in Fig. D-1.

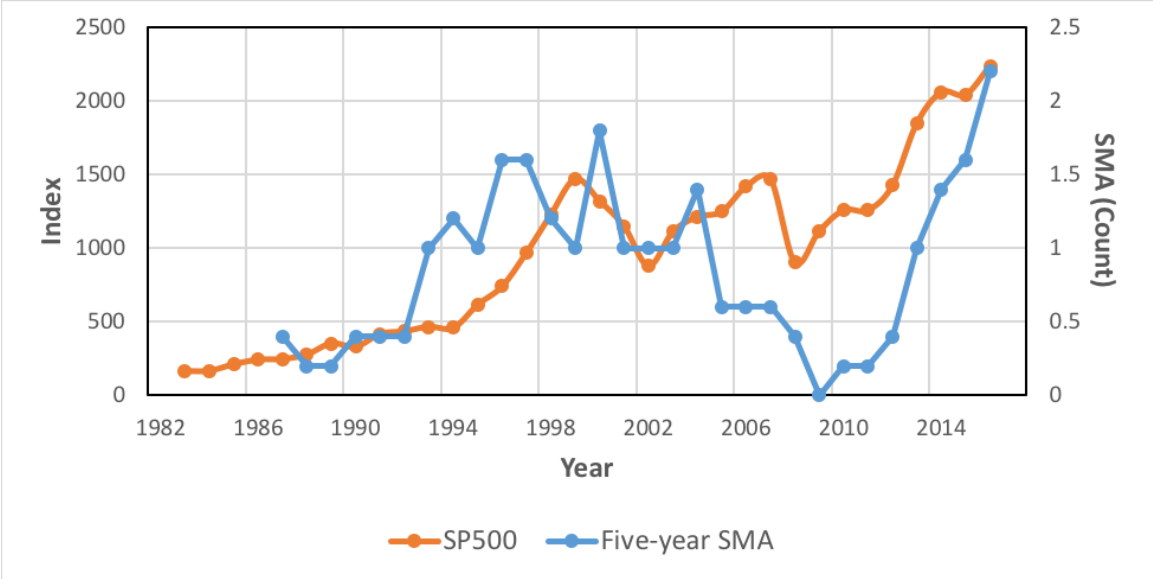


Figure D-1: Simple moving average of the number of MIT biotech companies that went through IPO/reverse-merger, plotted against the S&P 500 index. *Abbreviations*: SMA, simple moving average.

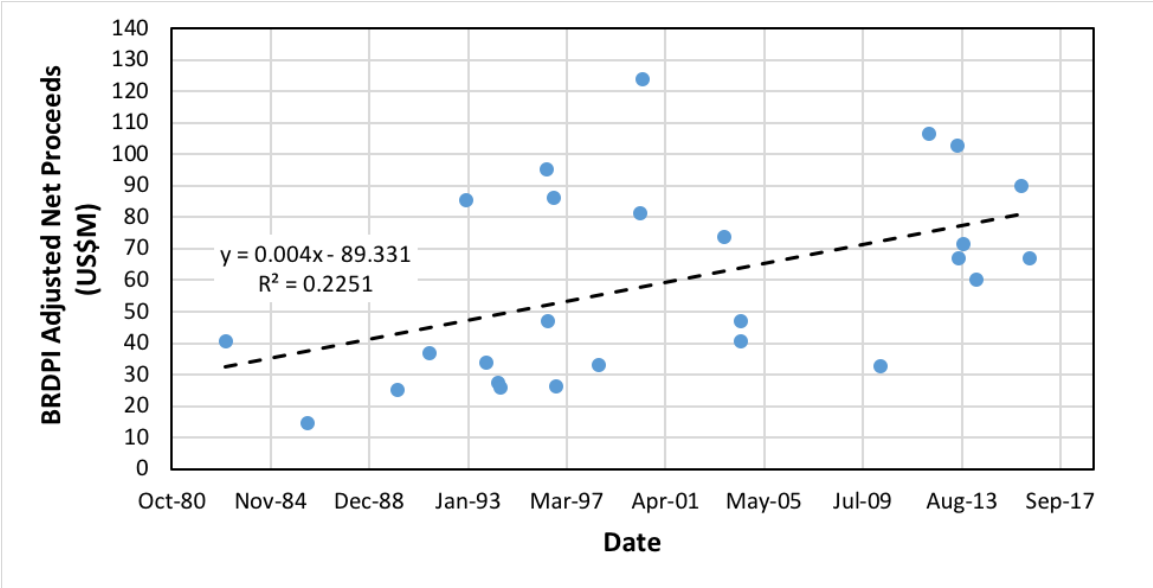


Figure D-2: BRDPI adjusted net proceeds for IPOs.

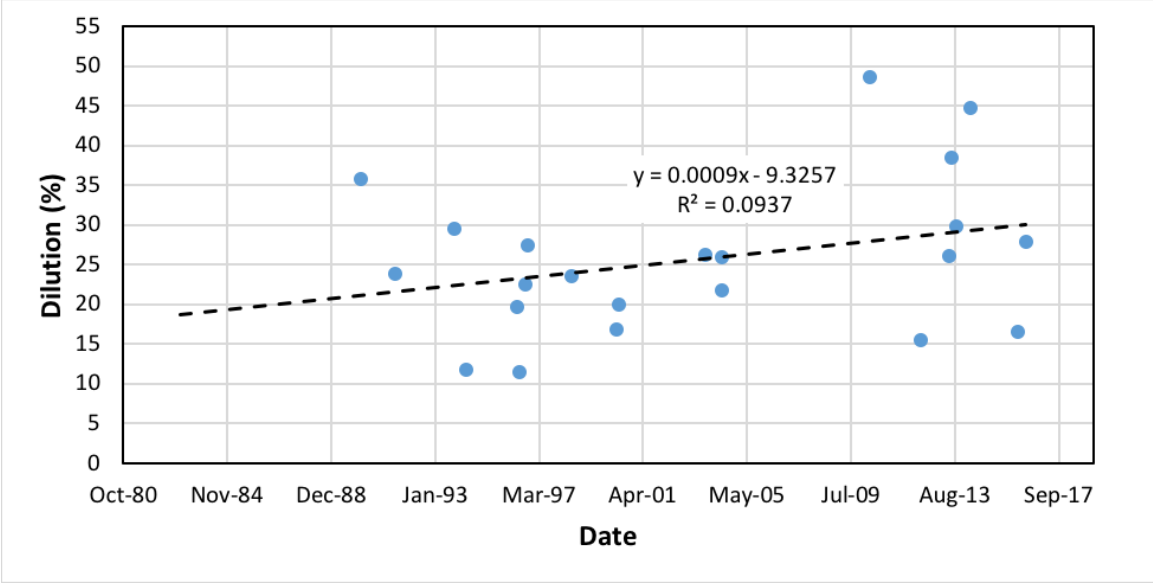


Figure D-3: IPO dilution.

D.4 Mergers and Acquisitions

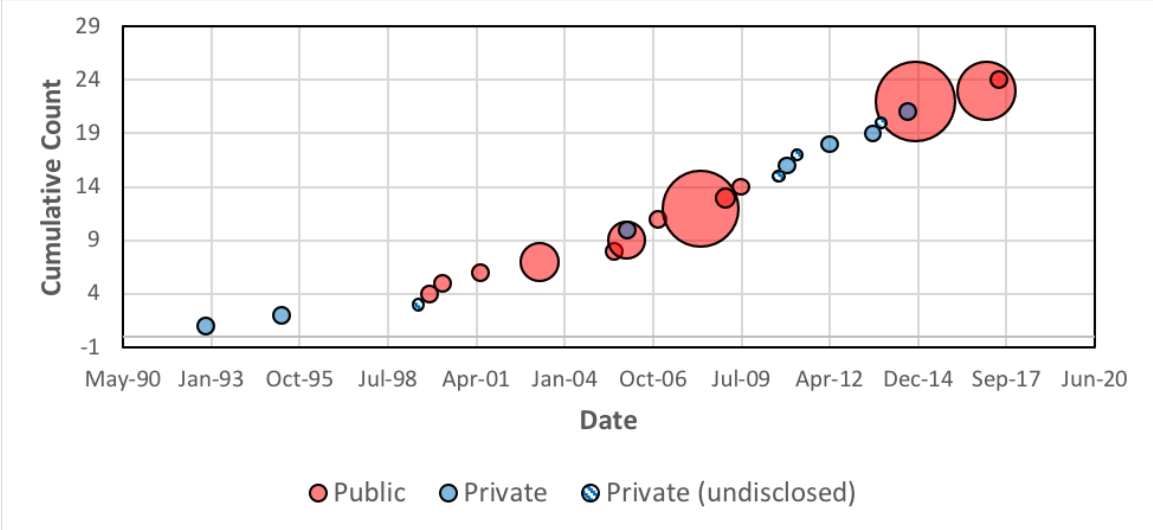


Figure D-4: Acquisition values of MIT biotech companies. The size of the marker indicates the size of the deal. The large M&A volume was driven primarily by three public company deals: Millennium, ARIAD, and Cubist.

D.5 Research and Development Pipeline

We propose to quantify the economic and biomedical impact of MIT licensees by measuring the quantity, the depth (the extent of development), and the breadth (the therapeutic areas involved) of candidates in each company's pipeline. In order to track their R&D pipelines, we went through each company's annual financial filings to the SEC (Form 10-K) throughout their entire lifespan and manually extracted all investigational compounds in the clinical phase. (The EDGAR archive contains filings only from 1994 onwards.)

In constructing our dataset, we chose to exclude preclinical candidates from consideration. This was done for several reasons. Companies are often vague in describing preclinical programs, perhaps for strategic purposes, typically describing only the broad direction of research without going into detail about the drugs under investigation. Furthermore, preclinical candidates are often excluded from financial filings altogether. Consequently, it is difficult to determine the exact number of such candidates accurately.

Our study focuses on the number of unique pipeline drug candidates (quantity), the highest stage of development for any indication (depth), and the therapeutic areas involved (breadth), with less emphasis given to the specific number of indications targeted. This is to maintain consistency with the approach of Mullard [151], which excludes label expansions, but it is also due to the difficulties in quantifying the scope of a candidate during its early clinical phases. It is not uncommon for companies to adopt the "shotgun" approach in early phases, enrolling patients with several different diseases that may be related, with the intention of converging to a lead indication in later phases. For example, BIND Therapeutics' BIND-014 in its phase 2 trial studied patients with urothelial carcinoma, cholangiocarcinoma, cervical cancer, and head and neck cancer, while Millennium's MLN341 phase 1 trial studied patients with a variety of solid tumors, including colon, breast, pancreatic, and prostate cancers. In these cases, the number of targeted indications is unclear.

Summary statistics of the dataset are shown in Tables D.2 and D.3, broken down

by company, stage of development, and indication group.

Table D.2: Pipeline candidates in 10-K filings by highest development stage. Editas Medicine and Enumeral Biomedical did not report any pipeline drugs in clinical trials in their 10-K. *Abbreviations*: n, number of unique drugs.

	Year				Highest Development Stage					
Company	**Start**	**End**	**Span**	**Status**	**P1**	**P2**	**P3**	**NDA**[1]	**APP**[2]	**n**
Scios	1994	2002	9	Acquired	1	3	1	1	1	7
Indevus	1995	2008	14	Acquired	5	1	6	2	7	21
Celldex	1995	2017	23	Alive	7	13	1		1	22
Algos	1996	1999	4	Acquired	2	5	1	1		9
Axys	1996	2000	5	Acquired		2				2
Cistron	1996	2000	5	Acquired		1				1
Guilford	1996	2004	9	Acquired	2	2	2		2	8
Millennium	1996	2007	12	Acquired	13	10	1		2	26
Cell Genesys	1996	2008	13	Acquired	1	3	1			5
Cubist	1996	2013	18	Acquired	3	3	2		6[A]	14
ARIAD	1996	2015	20	Acquired		2		1	1	4
Alkermes	1996	2017	22	Alive	8	8	5	3	22	46
Adhera	1996	2017	22	Alive	12	7		2	3	24
Abgenix	1998	2005	8	Acquired	2	1	1	1		5
Praecis	2000	2005	6	Acquired	1	2			1	4
Curis	2000	2017	18	Alive	4	3	1		1	9
Insmed	2000	2017	18	Alive	2	3	1	1	1	8
Sangamo	2000	2017	18	Alive	2	4				6
Acusphere	2003	2007	5	Bankrupt	2		1			3
Alnylam	2004	2017	14	Alive	4	7	4	1		16
Momenta	2004	2017	14	Alive	3	2	1	1	2	9
Avicena	2005	2007	3	Bankrupt		3	2			5
Noxxon[3]	2006	2017	12	Alive		3				3
Tengion	2010	2013	4	Bankrupt	2	1				3
Merrimack	2011	2017	7	Alive	2	4			1	7
BIND	2013	2015	3	Bankrupt	1	1				2
bluebird	2013	2017	5	Alive	1	1	2			4
Conatus	2013	2017	5	Alive		1				1
Enumeral	2013	2017	5	Alive						
Cerulean	2014	2017	4	Acquired		2				2
Editas	2015	2017	3	Alive						
Synlogic	2015	2017	3	Alive	1	1				2
Selecta	2016	2017	2	Alive	2	1				3
Total					83	100	33	14	51	281

[1] Includes BLA. [2] Includes drugs acquired post-approval and withdrawn products.
[3] Pipeline determined from Noxxon Pharma's press releases and publications.
[A] Zerbaxa and Sivextro reported approval in 10-Qs just prior to acquisition.

Table D.3: Pipeline candidates in 10-K filings by indication group. Editas Medicine and Enumeral Biomedical did not report any pipeline drugs in clinical trials in their 10-K. *Abbreviations*: cardio, cardiovascular; derm, dermatological; diges, digestive; genit, genitourinary; hema, hematological; horm, hormonal; immu, immunological; infec, infection; meta, metabolic; musc, musculoskeletal; neuro, neurological; onco, oncology; resp, respiratory.

	Indication Group												
Company	**Cardio**	**Derma**	**Digest**	**Genit**	**Hema**	**Horm**	**Immu**	**Infec**	**Meta**	**Musc**	**Neuro**	**Onco**	**Resp**
Scios	3	1		2			2		1	2	1		
Indevus	2		2	9		4	3	2	2		4	1	1
Celldex	2	1	2	4	2		3	7			2	12	2
Algos				1							9		
Axys		1	1				1						1
Cistron		1											
Guilford	1			2							5	2	1
Millennium	8		2	4	4		8		1	4	8	10	3
Cell Genesys		1	1	4	1		1	1				5	1
Cubist	1	3	4	1	1			10			1		
ARIAD				1	2		1			1	1	3	1
Alkermes	4	1	2	2		4	2		5	3	24	3	3
Adhera	1		2	1	2	2		1	4	2	10	5	
Abgenix		2	1	1		1	2			1		3	2
Praecis				1	1	1					2	2	
Curis	1	1	1	3	4							7	2
Insmed	1	1	1	3		3	1	2	3	3		3	5
Sangamo	1				2		1	1	1	1	3	1	
Acusphere													1
Alnylam	2			2	1			2	5		4	1	
Momenta	1	1	1		2		4			2	1	1	
Avicena									1	2	5		
Noxxon[1]			1	1	2				1			1	
Tengion				2						1			
Merrimack			2	1							1	7	2
BIND				1								2	1
bluebird					3				1		1	2	
Conatus	1							1	1				
Enumeral													
Cerulean			1	1								2	
Editas													
Synlogic				1	1				1		1	1	
Selecta										1	1	1	
Total	29	14	24	48	28	15	29	27	27	23	84	75	26

[1] Pipeline determined from Noxxon Pharma's press releases and publications.

D.6 Drug Approvals

Table D.4: List of FDA-approved drugs with MIT licensee ownership or contribution. Partner tag applies if another firm led clinical development of the drug enabled by the MIT company's technology. Originator tag applies if the drug was acquired by the MIT company. An asterisk indicates that the drug acquired was a post-phase 3 asset (NDA-ready, NDA-filed, or marketed). Acquirer tag applies if the MIT company was lead developer but was later acquired.

Company	Drug	Indication	Properties	Approved	NME/NBE	PR	Partner	Originator	Acquirer
Adhera	Stadol	Pain	Transnasal opioid	1991			BMS		
Adhera	Nascobal	Vitamin B12 deficiency	Intranasal cyanocobalamin	2005					Questcor
Adhera	Prestalia	Hypertension	ACE inhibitor, calcium channel blocker	2015				Symplmed*	
Alkermes	Verelan	Hypertension	Calcium channel blocker	1990			Cephalon	Elan*	
Alkermes	Cardizem CD	Hypertension, Angina	Calcium channel blocker	1991			Cephalon	Elan*	
Alkermes	Naprelan	Mild-to-moderate pain	COX-1, COX-2 antagonist	1996			Shionogi	Elan*	
Alkermes	Zanaflex	Muscle spasticity	$\alpha 2$ adrenergic agonist	1996	NME		Acorda	Elan*	
Alkermes	Rapamune	Renal transplant rejection	mTOR inhibitor	1999	NME	PR	Pfizer	Elan*	
Alkermes	Nutropin depot	Growth deficiency	rHGH [XR]	1999		PR	Genentech		
Alkermes	Afeditab CR	Hypertension	Calcium channel blocker	2001			Watson	Elan*	
Alkermes	Avinza	Chronic moderate to severe pain	Opioid [XR]	2002			Pfizer	Elan*	
Alkermes	Ritalin LA	ADHD	Methylphenidate [XR]	2002			Novartis	Elan*	
Alkermes	Emend	Chemo and surgery-associated nausea	NK1 antagonist	2003	NME	PR	Merck	Elan*	
Alkermes	Risperdal Consta	Schizophrenia	Risperidone [XR]	2003			Janssen		
Alkermes	Tricor 145	Cholesterol lowering	PPARa agonist	2004			Abbvie	Elan*	
Alkermes	Focalin XR	ADHD	Methylphenidate [XR]	2005			Novartis	Elan*	
Alkermes	Megace ES	Cachexia associated with AIDS	Varied	2005			Strativa	Elan*	
Alkermes	Vivitrol	Alcohol dependence	Naltrexone [XR]	2006		PR			
Alkermes	Luvox CR	Obsessive-compulsive disorder	$\sigma 1$ receptor agonist	2008			Jazz	Elan*	
Alkermes	Invega Sustenna	Schizophrenia	Paliperidone [XR]	2009			Janssen	Elan*	
Alkermes	Ampyra	Multiple sclerosis	Potassium channel blocker [XR]	2010	NME	PR	Acorda	Elan*	
Alkermes	Bydureon	Type 2 diabetes	GLP-1 agonist [XR]	2012			Astrazeneca		
Alkermes	Zohydro	Pain	Opioid [XR]	2013			Zogenix	Elan	
Alkermes	Aristada	Schizophrenia	Aripiprazole [XR]	2015	NME				
Alkermes	Invega Trinza	Schizophrenia	Paliperidone [XR]	2015		PR	Janssen	Elan	
ARIAD	Iclusig	T315+ CML & ALL, Ph+ ALL	BCR-ABL	2012	NME	PR			Takeda
Celldex	Rotarix	Rotavirus gastroenteritis	Rotavirus vaccine	2008	NBE		GSK	Avant*	
Cubist	Merrem	Broad-spectrum antibiotic	Carbapenem	1996	NME			Astrazeneca*	Pfizer
Cubist	Cubicin	Gram-positive cSSSI	Antibiotic	2003	NME	PR		Eli Lilly	Merck
Cubist	Entereg	Post-surgery GI recovery	Peripheral μ-opioid antagonist	2008	NME			Adolor*	Merck
Cubist	Dificid	C. diff-associated diarrhea	RNA polymerase-σ inhibitor	2011	NME	PR		Optimer*	Merck
Cubist	Sivextro	ABSSSI	Oxazolidinone	2014	NME	PR		Trius*	Merck
Cubist	Zerbaxa	Abdominal and urinary tract infections	Cephalosporin	2014	NME	PR		Calixa	Merck
Curis	Erivedge	Advanced BCC	Hedgehog signaling inhibitor	2012	NME	PR	Genentech		
Guilford	Gliadel	Glioblastoma multiforme	Chemotherapy-loaded wafer	1996		PR			Arbor
Guilford	Aggrastat	Acute coronary syndrome	Fibrinogen inhibitor	1998		PR		Merck*	
Indevus	Delatestryl	Hypogonadism	Testosterone	1953				Savient*	Endo
Indevus	Redux	Obesity	Serotonergic anorectic	1996	NME		Wyeth		
Indevus	Sanctura	Overactive bladder	Muscarinic antagonist	2004	NME			Madaus	Endo
Indevus	Vantas	Prostate cancer	GnRH agonist	2004				Valera*	Endo
Indevus	Supprelin LA	Precocious puberty	GnRH agonist	2007				Valera*	Endo
Indevus	Sanctura XR	Overactive bladder	Muscarinic antagonist [XR]	2007					Endo
Insmed	Iplex	Severe primary IGF-1 deficiency	Recombinant IGF-1 and IGFBP-3	2005	NBE	PR		Celtrix	
Merrimack	Onivyde	Metastatic pancreatic adenocarcinoma	Topoisomerase-1 inhibitor	2015		PR		Pharmengine	Ipsen
Millennium	Integrilin	Acute coronary syndrome	Inhibits platelet aggregation	1998	NME	PR		COR Therapeutics*	Schering
Millennium	Campath	Chronic lymphocytic leukemia	CD52 inhibitor (IgG1κ)	2001	NBE	PR		LeukoSite*	Genzyme
Millennium	Velcade	Multiple myeloma	Proteasome inhibitor	2003	NME	PR		LeukoSite	Takeda
Momenta	Enoxaparin Sodium	Acute heart attack	Lovenox generic	2014					
Momenta	Glatopa	Multiple sclerosis	Copaxone generic	2015					
Praecis	Plenaxis	Advanced prostate cancer	GnRH inhibitor	2003	NME	PR			
Scios	Natrecor	Acute congestive heart failure	B-type natriuretic peptide	2001	NME				

Table D.5: Label expansions into new disease indications by MIT licensees. We define a label expansion as FDA approval in a different disease than the first approval, omitting expansions into various lines of therapy or patient populations within a disease.

Company	Drug	Indication	Properties	Approved	NME/NBE	PR	Partner	Originator	Acquirer
Alkermes	Risperdal Consta	Bipolar disorder	Risperidone [XR]	2009			Janssen		
Alkermes	Vivitrol	Opioid dependence	Naltrexone [XR]	2010		PR			
Alkermes	Invega Sustenna	Schizoaffective disorder	Paliperidone [XR]	2014		PR	Janssen		
Cubist	Cubicin	S. aureus blood infections	Antibiotic	2006		PR		Eli Lilly	Merck
Indevus	Sarafem	Premenstrual dysphoric disorder	SSRI	2000			Lilly	Lilly	Endo
Millennium	Velcade	Mantle cell lymphoma	Proteasome inhibitor	2006		PR		LeukoSite	Takeda
Millennium	Campath (Lemtrada)	Multiple sclerosis	CD52 inhibitor (IgG1κ)	2014				LeukoSite	Genzyme[1]

[1] Genzyme acquired Millennium's equity stake in Campath in 2004.

Table D.6: Post-acquisition drug approvals by MIT licensees.

Company	Drug	Indication	Properties	Approved	NME/NBE	PR	Partner	Originator	Acquirer
Abgenix	Vectibix	EGFR+ mCRC	EGFR+ inhibitor (IgG2)	2006	NBE	PR	Amgen		Amgen
Abgenix	Prolia	High-risk osteoporosis	RANKL inhibitor (IgG2)	2010	NBE	PR	Amgen		Amgen
Abgenix	Repatha	LDL-C lowering	PCSK9 inhibitor (IgG2)	2015	NBE		Amgen		Amgen
Abgenix	Imfinzi	Advanced UC	PD-L1 inhibitor (IgG1κ)	2017	NBE	PR	Astrazeneca		
ARIAD	Alunbrig	ALK+ NSCLC	ALK, EGFR inhibitor	2017	NME	PR	Takeda		Takeda
Indevus	Aveed	Male hypogonadism	Testosterone prodrug	2014			Endo	Schering AG	Endo
Millennium	Ninlaro	Multiple myeloma	Proteasome inhibitor	2015	NME	PR	Takeda	Leukosite	Takeda

D.7 Intellectual Property

The USPTO database was used to collect data for the 33 MIT licensees. Using the patent assignee name, the database was queried for the patent number and date of granting. We limit the dataset to patents granted rather than to patents filed because not all patents are ultimately approved. Number of patents per year for a given company was calculated by dividing the total patents by the lifetime of the company since inception. The dataset cutoff is December 31, 2017.

Company specific licensing data, patent data, summary statistics, and cumulative patent data for the MIT portfolio of companies are shown in Table D.7 and Fig. D-5. The 33 companies licensed 258 unique patents from MIT in initial startup agreements totaling $39.9M in royalties. These companies were additionally granted 2,512 patents between 1985 and 2017, clearly showing that MIT licensees continue to innovate beyond an initial license from MIT. The trend over time in number of yearly patents granted roughly mimics the cycles in the public markets. This pattern is likely due to the fact that capital availability fueled significant R&D investments. For example, the 1999 technology bubble provided significant sums of capital to biotechnology firms, such as Millennium Pharmaceuticals. Millennium raised over $1B in 2000 alone from public market investors and also achieved 202 patents granted from 2000–2002.

The most "productive" companies, defined by the number of patents granted per year, are shown in Table D.7. As might be expected, these companies—including Millennium Pharmaceuticals, Alnylam, and Sangamo Biosciences—developed novel technology platforms. Millennium developed technology using new genomic insights and technology; Alnylam pioneered the cutting-edge siRNA technology discovered by its co-founder Phillip Sharp; and Sangamo Biosciences similarly worked on its unique zinc-finger nuclease gene-editing platform. In contrast, companies focused on developing in-licensed assets, such as Conatus Pharmaceuticals, tend to be focused on clinical development rather than on innovating new drug discovery and development technology. Conatus acquired emricasan, a compound for liver disease, from Pfizer, who itself acquired the drug from Idun, co-founded by MIT's Robert Horvitz [223].

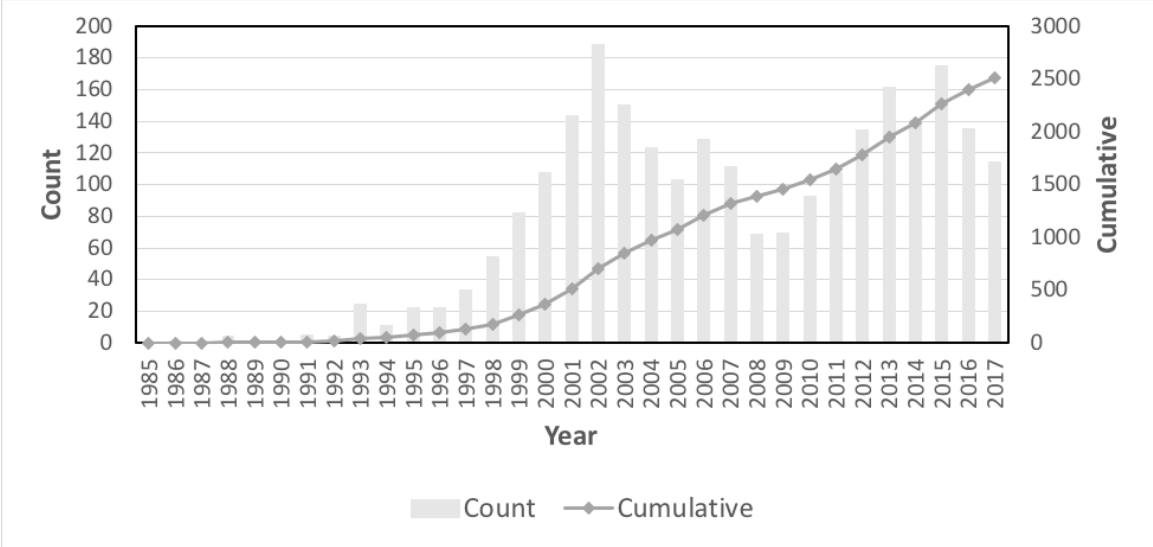


Figure D-5: Cumulative patents granted to MIT licensees from 1985–2017.

Fig. D-6 correlates the number of drug approvals (excluding Originators*) versus the number of patents granted per year for MIT licensees. The correlation is complicated by the blend of companies in the MIT dataset. Some developed novel technology platforms, while others focused on acquiring and developing assets, which inherently do not rely on securing new patents. If one considers the number of patents to be a proxy for the strength of a drug development technology platform, the poor correlation supports the prior conclusion that many of the drugs under ownership by MIT licensees were acquired from Originators—drug approvals from originators have little relation to the company's own drug discovery capabilities. The lack of correlation also highlights that developing new drugs is a difficult business, as new technologies and discoveries do not necessarily translate to new products.

Table D.7: Patents licensed by and granted to MIT licensees from 1985–2017.

Company	Status	Licensed MIT Patents	Granted Patents	Patents Granted/Year
Abgenix[1]	Acquired			
Algos	Acquired	1	9	2
ARIAD	Acquired	3	69	3
Axys	Acquired	1	42	4
Cell Genesys	Acquired	16	84	4
Cerulean[2]	Acquired			
Cistron	Acquired	9	2	
Cubist	Acquired	2	64	3
Guilford	Acquired	29	104	9
Indevus	Acquired	20	29	2
Millennium	Acquired	1	803	50
Praecis[2]	Acquired			
Scios	Acquired	18	127	8
Adhera[2]	Alive			
Alkermes	Alive	1	212	10
Alnylam	Alive	9	232	15
bluebird	Alive	7	7	
Celldex	Alive	7	48	2
Conatus	Alive	6		
Curis	Alive	28	118	6
Editas	Alive	10	1	
Enumeral	Alive	10		
Insmed	Alive	6	32	1
Merrimack[2]	Alive			
Momenta	Alive	39	78	5
Noxxon	Alive	1	23	1
Sangamo	Alive	10	155	7
Selecta	Alive	23	10	1
Synlogic	Alive	7	2	
Acusphere	Bankrupt	6	27	2
Avicena	Bankrupt	1	5	
BIND	Bankrupt	14	39	4
Tengion	Bankrupt	6	4	
Total (unique)		258	2512	

[1] Cell Genesys spin-off, may have used MIT IP through the license to Cell Genesys.
[2] Licensed MIT IP for which no US patent applications were issued (abandoned prior to issuance).

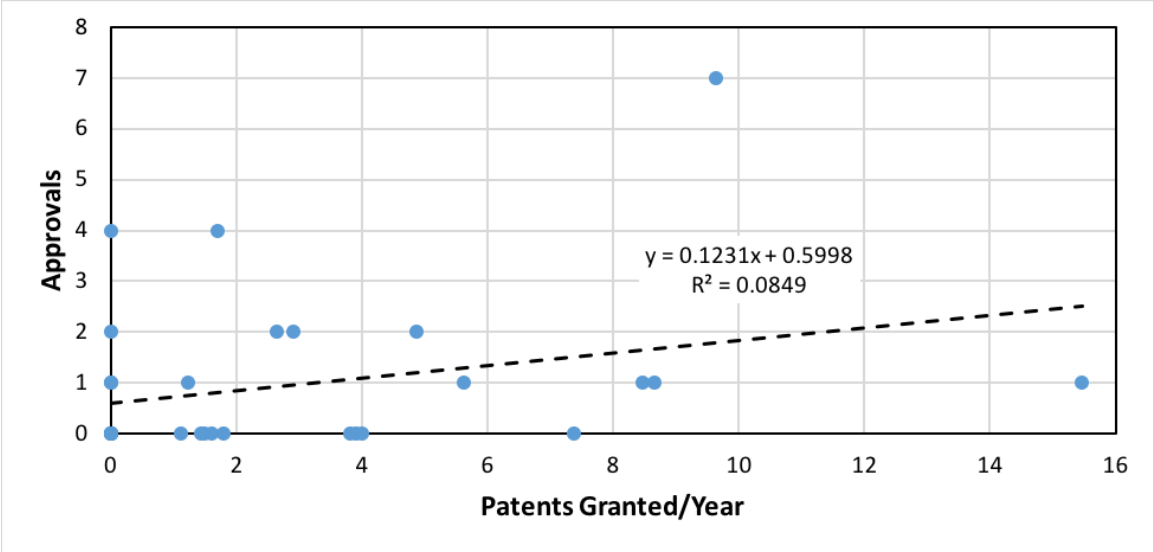


Figure D-6: Correlation between number of drug approvals and number of patents granted per year for MIT licensees. We exclude Millennium, an outlier with 50 patents granted per year on average and two approvals, from the plot to obtain a better regression fit.

D.8 FDA Approvals

Table D.8: FDA NDA/BLA and NME/ NBE drug approvals by year and by review priority between 1991–2017 [152, 175, 224, 225].

	NDA/BLA				NME/NBE			
Year	**Priority**	**Standard**	**Total**	**Priority-Total Ratio (%)**	**Priority**	**Standard**	**Total**	**Novel-Total Ratio (%)**
1991	19	44	63	30	14	16	30	48
1992	17	74	91	19	11	15	26	29
1993	19	51	70	27	13	12	25	36
1994	16	45	61	26	12	9	21	34
1995	16	67	83	19	10	19	29	35
1996	29	102	131	22	18	35	53	40
1997	20	101	121	17	9	30	39	32
1998	25	65	90	28	16	14	30	33
1999	28	55	83	34	19	16	35	42
2000	20	78	98	20	9	18	27	28
2001	10	56	66	15	7	17	24	36
2002	11	67	78	14	7	10	17	22
2003	14	58	72	19	9	12	21	29
2004[A]	29	89	118	25	21	15	36	31
2005	22	58	80	28	15	5	20	25
2006	21	80	101	21	10	12	22	22
2007	23	55	78	29	8	10	18	23
2008	18	70	88	20	9	15	24	27
2009	19	78	97	20	8	18	26	27
2010	16	76	92	17	11	10	21	23
2011	26	68	94	28	15	15	30	32
2012	23	77	100	23	16	23	39	39
2013	16	82	98	16	9	18	27	28
2014	34	84	118	29	24	17	41	35
2015	37	84	121	31	24	21	45	37
2016	24	70	94	26	15	7	22	23
2017	47	96	143	33	28	18	46	32
Total	599	1,930	2,529	24	367	427	794	31

[A] Beginning in Fiscal Year 2004, Center for Drug Evaluation and Research (CDER) started reviewing therapeutic biologic products transferred from Center for Biologics Evaluation and Research (CBER) to CDER.

D.9 Additional Discussion

As shown by the analysis of IP generated by the MIT licensees, entirely new classes of potentially transformative drugs are on the horizon whose development uses MIT IP. A brief survey of drugs in the development phase underscores MIT's contribution to the current creation of innovative therapeutics. Alnylam, bluebird bio, Editas Medicine, and Sangamo are four examples of firms that have licensed MIT IP to develop platform technologies capable of discovering new drugs.

Alnylam was founded in 2002 by a group of MIT faculty, including Robert Langer and Phillip Sharp, to develop RNAi therapeutics based on siRNA discoveries. The firm licensed patents from MIT on the formulation and delivery of siRNAs, and also engaged in a five-year research collaboration to improve delivery to target tissues. Alnylam has a broad clinical development pipeline of seven assets, including patisiran, an RNAi therapeutic for hereditary ATTR amyloidosis, a severe neuropathy, under priority review by the FDA as of the cutoff of our dataset [185]. (In August 2018, Alnylam received FDA approval for patisiran as a first-of-its kind targeted RNA-based therapy [179].)

One of bluebird bio's lead candidates is LentiGlobin, a lentiviral-based gene therapy for transfusion-dependent β-thalassemia, an inherited blood disorder. LentiGlobin is currently in a phase 3 clinical trial. The company co-owns a patent portfolio with MIT containing patents from Irving London's laboratory on the specific composition of lentiviral β-globin expression vectors.

Sangamo is pioneering the use of zinc finger proteins (ZFPs) in gene editing, a technology with its roots at MIT in Carl Pabo's lab [226, 227]. In 1996, MIT granted Sangamo an exclusive license to its ZFP IP, although the company has not yet developed a drug to a market. The firm has five drug candidates in phase 1/2 trials enabled by ZFP technology for hemophilia B, MPS (mucopolysaccharidosis) I, MPS II, and HIV.

Most recently, Editas was formed in 2013 to commercialize CRISPR/Cas9 gene editing technology, in part invented by Feng Zhang at the Broad Institute of MIT and

Harvard. Though Editas was founded too recently to have developed an extensive patent portfolio or any clinical-stage drugs, its platform technology has the potential to transform the treatment paradigm for a vast array of diseases.

Appendix E

Supplement to Chapter 6

E.1 S&P Historical Default Rates

Table E.1: Global corporate average cumulative default rates by rating modifier (1981–2018). Sources: S&P Global Fixed Income Research and S&P Global Market Intelligence's CreditPro®.

(%)	Time horizon (years)														
Rating	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**
AAA	0.00	0.03	0.13	0.24	0.35	0.45	0.51	0.59	0.65	0.70	0.73	0.76	0.79	0.85	0.92
AA+	0.00	0.05	0.05	0.10	0.15	0.21	0.26	0.32	0.38	0.44	0.50	0.56	0.62	0.69	0.76
AA	0.02	0.03	0.08	0.22	0.36	0.48	0.60	0.71	0.80	0.89	0.97	1.03	1.14	1.20	1.27
AA-	0.03	0.08	0.17	0.25	0.32	0.44	0.50	0.55	0.61	0.66	0.72	0.79	0.81	0.85	0.90
A+	0.05	0.09	0.20	0.33	0.43	0.53	0.64	0.76	0.89	1.03	1.17	1.31	1.47	1.66	1.83
A	0.06	0.14	0.23	0.35	0.48	0.65	0.83	1.00	1.19	1.41	1.59	1.73	1.86	1.95	2.12
A-	0.06	0.16	0.26	0.38	0.54	0.70	0.93	1.10	1.23	1.34	1.45	1.58	1.71	1.83	1.92
BBB+	0.10	0.29	0.50	0.73	0.97	1.25	1.46	1.68	1.93	2.17	2.41	2.58	2.80	3.07	3.37
BBB	0.16	0.41	0.64	1.01	1.36	1.72	2.04	2.36	2.72	3.08	3.46	3.77	4.01	4.12	4.33
BBB-	0.24	0.73	1.35	2.04	2.77	3.42	4.00	4.55	5.00	5.39	5.83	6.19	6.51	7.00	7.37
BB+	0.32	1.04	1.91	2.79	3.69	4.56	5.29	5.81	6.42	7.04	7.45	7.95	8.43	8.77	9.27
BB	0.53	1.61	3.19	4.68	6.17	7.35	8.43	9.35	10.22	10.98	11.76	12.39	12.81	13.12	13.53
BB-	0.95	2.98	5.11	7.33	9.27	11.15	12.71	14.21	15.42	16.46	17.28	17.99	18.74	19.48	20.15
B+	2.01	5.52	8.95	11.88	14.15	15.89	17.54	18.97	20.30	21.49	22.48	23.14	23.80	24.45	25.09
B	3.41	7.84	11.69	14.73	17.09	19.27	20.74	21.77	22.74	23.74	24.48	25.18	25.77	26.30	26.85
B-	6.75	13.73	19.04	22.70	25.43	27.42	29.01	30.11	30.82	31.37	32.13	32.67	32.91	33.18	33.50
CCC/C	26.89	36.27	41.13	43.94	46.06	46.99	48.20	49.04	49.80	50.44	50.96	51.51	52.16	52.72	52.80
Investment grade	0.09	0.25	0.43	0.66	0.90	1.14	1.36	1.56	1.77	1.96	2.16	2.32	2.48	2.63	2.80
Speculative grade	3.66	7.13	10.12	12.56	14.55	16.18	17.55	18.69	19.70	20.62	21.39	22.02	22.60	23.13	23.65
All rated	1.48	2.91	4.16	5.21	6.08	6.82	7.44	7.97	8.44	8.88	9.26	9.58	9.87	10.13	10.41

Appendix F

Supplement to Chapter 7

F.1 Literature Estimates

Table F.1: Literature estimates of probability of success, costs of development, and duration at each phase of development for standard clinical trials. *Abbreviations*: PoS, probability of success; PRE, preclinical; P1, phase 1; P2, phase 2; P3, phase 3; NDA, New Drug Application.

Parameter	PRE to P1	P1 to P2	P2 to P3	P3 to NDA	NDA to Approval	PRE to Approval
PoS (%)	69.0[A]	81.4[B]	30.5[B]	26.5[B]	100.0	4.5
Duration (months)	12..0[C]	42.1[D]	40.6[D]	48.5[D]	9.6[E]	152.8
Development cost ($ millions)	1.2[C]	10.1[F]	20.7[F]	92.8[F]	0.0[F]	127.1
Upfront cost ($ millions)	2.3[G]	0.0	0.0	0.0	0.0	
Discount factor (%)	23.0[H]	20.0[H]	17.2[H]	12.5[H]	10.0[H]	

[A] For new molecular entities (NMEs) [209]. [B] For GBM [208].
[C] For small molecule oncology NMEs [210], assuming acquisition after lead optimization.
[D] Median for glioblastoma [208]. [E] For orphan drugs [211]. [F] Assume development cost is similar to that for orphan drugs since most brain cancers are rare diseases [189], adjusted to 2019 dollars using the Biomedical Research and Development Price Index (BRDPI)
[G] For small molecule oncology NMEs [210], assume an upfront cost that is equal to the cost required to complete lead optimization, adjusted to 2019 dollars using BRDPI.
[H] Assume costs of capital are similar for oncology drugs [202].

F.2 Estimates by Experts

Table F.2: Probability of success, costs of development, and duration at each phase of development for standard clinical trials as estimated by the experts at NBTS. To reduce the impact of outliers, we use the median of the estimates provided by the panel. *Abbreviations*: PoS, probability of success; PRE, preclinical; P1, phase 1; P2, phase 2; P3, phase 3; NDA, New Drug Application.

Parameter	**PRE to P1**	**P1 to P2**	**P2 to P3**	**P3 to NDA**	**NDA to Approval**	**PRE to Approval**
PoS (%)	60.0	60.0	40.0	45.0	100.0	6.5
Duration (months)		24.0	38.0	51.5	12.0	
Development cost ($ millions)	1.0	6.5	16.5	70.0		

F.3 Cost and Duration of GBM AGILE

We derive cost and duration estimates for GBM AGILE assuming a steady state of four arms (one control and three experimental arms), an accrual rate of 30 patients per month (taking into account the number of sites launched in the U.S., Canada, China, and Europe), and a cost per patient of $84,000 (estimates as of June 2020). In GBM AGILE, the cost of the common control arm is shared among the three experimental arms, i.e., each experimental arm incurs a cost of $28,000 per patient in the control. For simplicity, we allocate 20% of the newly enrolled patients each month to the control arm, and assign the remaining 80% evenly among the experimental arms. (In reality, the proportion allocated to each experimental arm will change over time according to its demonstrated efficacy, and is determined through Bayesian adaptive randomization.)

We assume that 100 patients are required for early graduation from stage 1 of GBM AGILE to stage 2, and 150 patients are required for regular graduation to stage 2, or a transition to a phase 2 or phase 3 trial. Due to the use of Bayesian adaptive randomization, we expect experimental arms that do not demonstrate efficacy to be allocated fewer patients over time before being discontinued. Therefore, we assume a smaller accrual of 50 patients for arms that are stopped for futility in stage 1. For stage 2, we assume that 50 patients are required for confirmation in a subgroup comprising 30% of the patient population (e.g., the newly diagnosed unmethylated, newly diagnosed methylated, or recurrent disease with additional stratification/enrichment biomarkers subgroups).

Given the rates of accrual and enrollment, the duration of an experimental arm is given by:

$$d = \frac{nm}{ves} + f \tag{F.1}$$

where d is the duration in months, n is the trial accrual required for graduation, transition, or futility (e.g., 100 patients for early graduation from stage 1), v is the overall monthly accrual rate (e.g., 30 patients per month), e is the proportion of

newly enrolled patients allocated to experimental arms (e.g., 80%), m is the number of experimental arms in steady state, s is the prevalence of the patient subtype under investigation (e.g., 30% for confirmation in stage 2), and f is the time added to allow for follow-up and data analysis after the last patient has been enrolled. The terms s and f are relevant only for stage 2; we assume that stage 1 encompasses all patient subtypes and the prevalence is 100%. Because GBM AGILE is designed to be a seamless platform trial, we assume that no follow-up time is required at the end of stage 1. On the other hand, we factor in an analysis and follow-up period of 18 months for stage 2.

We assume quarterly and semiannual payments for patient costs of the experimental arm and the control arm, respectively. They are given by:

$$c_e = pn \tag{F.2}$$

$$c_c = \frac{p}{m} \cdot n \cdot (1 - e) \tag{F.3}$$

where c_e is the cost due for the experimental arm in millions, c_c is the cost for the control arm, and p is the price per patient (e.g., \$84,000). Note that the cost per patient of the common control arm is divided among the m experimental arms. Furthermore, we assume that the number of control patients enrolled is limited to $(1 - e)$ of the experimental arm accrual (e.g., 20%). In addition to patient costs, we assume that an initiation fee of \$1.75M is due at the start of stage 1, an extension fee of \$1.5M at the start of stage 2, and a final fee of \$1.5M for data analysis at the end of stage 2. For our simulation, we discount these periodic cash flows to an equivalent single payment due at the start of each stage using a cost of capital of 15%.

F.4 Correlation

We first average the correlation estimates made by the experts. Next, we symmetrize the resulting correlation matrix by performing the following operation:

$$R = \frac{1}{2}(X + X^{\mathrm{T}}) \tag{F.4}$$

where X is the correlation matrix created from averaging the estimates by the experts, and R is a symmetric correlation matrix. Finally, we project the symmetric correlation matrix R to its nearest positive-definite counterpart Σ for use in our simulations [213].

To generate correlated trial outcomes, we first draw a vector of random multivariate standard normal variables $\epsilon_j \equiv [\epsilon_{1j}, \epsilon_{2j}, \ldots, \epsilon_{nj}]^{\mathrm{T}}$, where n is the number of projects in the portfolio, and j is the phase of development that is of interest. Next, we obtain $z_j \in \mathbb{R}^{n\times 1}$ by pre-multiplying ϵ_j with $\Sigma^{1/2}$, where $\Sigma^{1/2}$ denotes the Cholesky decomposition of Σ, a positive-definite matrix. The resulting vector z_j is consequently multivariate normal with covariance matrix Σ.

Given the probabilities of success in Tables 7.2 and 7.3, we can model trial outcomes as Bernoulli variables:

$$B_{ij} = \begin{cases} \text{Success}, & z_{ij} > \alpha_j \\ \text{Failure}, & z_{ij} \leq \alpha_j \end{cases} \tag{F.5}$$

where B_{ij} is the outcome for trial i in phase j, z_{ij} is component i of z_j, $\alpha_j = \Phi^{-1}(1 - p_j)$, Φ^{-1} is the inverse cumulative distribution function of a univariate standard normal distribution, and p_j is the probability of success for phase j.

F.5 Profitability of a Successful Compound

Table F.3: Assumptions for profitability of an approved drug for GBM. *Abbreviations*: HGGs, high-grade gliomas; uMGMT, unmethylated O6-methylguanine DNA methyltransferase; DRD2, dopamine receptor D2; EGFR, epidermal growth factor receptor.

	Market size per year	**Market penetration (%)**	**Price per patient ($)**
Newly diagnosed GBM and HGGs	16,500	10.0	66,000
Recurrent GBM	30,000	10.0	66,000
Newly diagnosed uMGMT GBM	9,900	20.0	66,000
Recurrent GBM with EGFR-low and DRD2-high tumor phenotype	9,000	20.0	66,000
Newly diagnozed GBM with EGFR	9,900	20.0	66,000
Pediatric gliomas	3,500	20.0	66,000
Brain metastases	70,000	10.0	66,000
Brain tumor	100,000	10.0	33,000

Table F.3 (continued): Assumptions for profitability of an approved drug for GBM.

	Value
Premium for transformative treatments	2.0x
Marketing exclusivity (years)	7.0
Pediatric extension (years)	0.5
Cost of capital (%)	10.0
Priority Review Voucher ($ millions)	100.0

Table F.4: NPV of projects on approval. *Abbreviations*: IMM, immunotherapy; DDR, DNA damage repair; TM, tumor metabolism; PM, precision medicine; DE, devices; DNA-PK, DNA-dependent protein kinase; ATM, ataxia-telangiectasia mutated; ATR, ataxia telangiectasia and Rad3-related protein; FGFR, fibroblast growth factor receptor; LPCAT1, lysophosphatidylcholine acyltransferase 1; DRD2, dopamine receptor D2; BBB, blood-brain barrier.

Therapeutic area	**Project**	**Market size per year**	**NPV ($ millions)**
IMM	T cell activation	30,000	1,928
	T cell activation	30,000	985
	T cell activation	30,000	1,928
	Personalized dendritic cell vaccine	16,500	1,214
	Retroviral replicating vectors	16,500	1,214
	Oncolytic virus	30,000	1,928
	Autologous tumor cell vaccine	16,500	1,060
DDR	DNA-PK inhibitor	9,900	1,272
	ATM inhibitor	9,900	1,272
	ATR inhibitor	16,500	1,060
	FGFR inhibitor	30,000	1,928
	DNA repair inhibitors	9,900	636
	ATM inhibitor	3,500	572
TM	LPCAT1 inhibitor	46,500	1,494
PM	DRD2 receptor antagonist	9,000	1,315
	BBB-penetrant signaling inhibitor	16,500	530
	CRISPR-Cas9 gene editing	46,500	2,988
	BBB-penetrant transcription factor inhibitor	16,500	530
	BBB-penetrant transcription factor inhibitor	70,000	2,249
DE	Fluorescence-guided surgery	100,000	1,607

Bibliography

[1] Jack W Scannell, Alex Blanckley, Helen Boldon, and Brian Warrington. Diagnosing the decline in pharmaceutical r&d efficiency. *Nature reviews Drug discovery*, 11(3):191–200, 2012.

[2] Deloitte. Ten years on. measuring the return from pharmaceutical innovation 2019. Available at https://www2.deloitte.com/uk/en/pages/life-sciences-and-healthcare/articles/measuring-return-from-pharmaceutical-innovation.html (Retrieved 2020-09-01), 2020.

[3] Graham Scholefield and Markus Thunecke. Global biopharma r&d productivity and growth rankings. Available at https://catenion.com/wp-content/uploads/2019/01/In-Vivo-Biopharma-RD-Productivity-Growth-2018.pdf (Retrieved 2020-09-01), 2019.

[4] MS Ringel, JW Scannell, M Baedeker, and U Schulze. Breaking eroom's law. *Nature reviews. Drug Discovery*, 2020.

[5] Andrew L Hopkins and Colin R Groom. The druggable genome. *Nature reviews Drug discovery*, 1(9):727–730, 2002.

[6] Fabio Pammolli, Laura Magazzini, and Massimo Riccaboni. The productivity crisis in pharmaceutical r&d. *Nature reviews Drug discovery*, 10(6):428–438, 2011.

[7] Bruce Booth and Rodney Zemmel. Prospects for productivity. *Nature Reviews Drug Discovery*, 3(5):451–456, 2004.

[8] Alexander Gaffney. It's not just you: Fda regulatory requirements really are increasing. Available at https://www.raps.org/regulatory-focus%E2%84%A2/news-articles/2014/10/it-s-not-just-you-fda-regulatory-requirements-really-are-increasing (Retrieved 2020-09-01), 2014.

[9] National Venture Capital Association. Nvca 2017 yearbook. Available at https://nvca.org/nvca-2017-yearbook-go-resource-venture-ecosystem/ (Retrieved 2020-09-01), 2017.

[10] National Venture Capital Association. Nvca 2020 yearbook. Available at https://nvca.org/wp-content/uploads/2020/03/NVCA-2020-Yearbook.pdf (Retrieved 2020-09-01), 2020.

[11] Stacy Lawrence. Biotech's wellspring—a survey of the health of the private sector in 2016. *Nature Biotechnology*, 35(5):413–420, 2017.

[12] Amy Brown, Elizabeth Cairns, and Edwin Elmhirst. Pharma, biotech & medtech 2019 in review. Available at https://www.evaluate.com/thought-leadership/vantage/evaluate-vantage-pharma-biotech-medtech-2019-review (Retrieved 2020-09-01), 2020.

[13] Kien Wei Siah, Sean Khozin, Chi Heem Wong, and Andrew W Lo. Machine-learning and stochastic tumor growth models for predicting outcomes in patients with advanced non–small-cell lung cancer. *JCO clinical cancer informatics*, 1:1–11, 2019.

[14] Andrew W Lo, Kien Wei Siah, and Chi Heem Wong. Machine learning with statistical imputation for predicting drug approvals. *Harvard Data Science Review*, July 2019.

[15] Andrew W Lo and Kien Wei Siah. Financing correlated drug development projects. *Journal of Structured Finance*, 2020.

[16] Richard T Thakor, Nicholas Anaya, Yuwei Zhang, Christian Vilanilam, Kien Wei Siah, Chi Heem Wong, and Andrew W Lo. Just how good an investment is the biopharmaceutical sector? *Nature biotechnology*, 35(12):1149–1157, 2017.

[17] Chi Heem Wong, Kien Wei Siah, and Andrew W Lo. Estimation of clinical trial success rates and related parameters. *Biostatistics*, 20(2):273–286, 2019.

[18] Andrew W Lo, Kien Wei Siah, and Chi Heem Wong. Estimating probabilities of success of vaccine and other anti-infective therapeutic development programs. *Harvard Data Science Review*, May 2020.

[19] Siah Kien Wei, Q Xu, K Tanner, O Futer, J J Frishkopf, and Andrew W Lo. Accelerating therapeutic innovation in glioblastoma treatments via nbts venture fund. Manuscript pending submission, 2020.

[20] S Huang, Kien Wei Siah, D Vasileva, S Chen, L Nelsen, and Andrew W Lo. The impact of university intellectual property: An analysis of mit life sciences technology licensing. Manuscript under review, 2019.

[21] D Berry, S Berry, P Hale, L Isakov, Andrew W Lo, Kien Wei Siah, and Chi Heem Wong. A cost/benefit analysis of clinical trial designs for covid-19 vaccine candidates. Manuscript under review, 2020.

[22] Scott Dessain and Scott E Fishman. *Preserving the Promise: Improving the Culture of Biotech Investment*. Academic Press, 2016.

[23] Jose-Maria Fernandez, Roger M Stein, and Andrew W Lo. Commercializing biomedical research through securitization techniques. *Nature biotechnology*, 30(10):964–975, 2012.

[24] Brady Huggett. Biotech's wellspring—a survey of the health of the private sector in 2014. *Nature biotechnology*, 33(5):470–477, 2015.

[25] John Hull, Andrew W Lo, and Roger M. Stein. Funding Long Shots. *Journal of Investment Management*, 17(4):1–33, 2019.

[26] Securities Industry and Financial Markets Association. Us fixed income issuance and outstanding. Available at https://www.sifma.org/resources/research/us-fixed-income-issuance-and-outstanding/ (Retrieved 2020-09-01), 2020.

[27] Andrew W Lo. Reading about the financial crisis: A twenty-one-book review. *Journal of economic literature*, 50(1):151–78, 2012.

[28] Daniel Morgensztern, Shean Huey Ng, Feng Gao, and Ramaswamy Govindan. Trends in stage distribution for patients with non-small cell lung cancer: A national cancer database survey. *Journal of Thoracic Oncology*, 5(1):29–33, Jan 2010.

[29] Alyson L. Mahar, Carolyn Compton, Lisa M. McShane, Susan Halabi, Hisao Asamura, Ramon Rami-Porta, and Patti A. Groome. Refining prognosis in lung cancer: A report on the quality and relevance of clinical prognostic tools. *Journal of Thoracic Oncology*, 10(11):1576–1589, Nov 2015.

[30] Larissa A Pikor, Varune R Ramnarine, Stephen Lam, and Wan L Lam. Genetic alterations defining NSCLC subtypes and their therapeutic implications. *Lung Cancer*, 82(2):179–189, 11 2013.

[31] Tsuyoshi Takahashi, Makoto Sonobe, Masashi Kobayashi, Akihiko Yoshizawa, Toshi Menju, Ei Nakayama, Nobuya Mino, Shotaro Iwakiri, Kiyoshi Sato, Ryo Miyahara, Kenichi Okubo, Toshiaki Manabe, and Hiroshi Date. Clinicopathologic features of non-small-cell lung cancer with EML4-ALK fusion gene. *Annals of Surgical Oncology*, 17(3):889–897, 3 2010.

[32] E. A. Eisenhauer, P. Therasse, J. Bogaerts, L. H. Schwartz, D. Sargent, R. Ford, J. Dancey, S. Arbuck, S. Gwyther, M. Mooney, L. Rubinstein, L. Shankar, L. Dodd, R. Kaplan, D. Lacombe, and J. Verweij. New response evaluation criteria in solid tumours: Revised RECIST guideline (version 1.1). *European Journal of Cancer*, 45(2):228–247, 1 2009.

[33] BioOncology. Clinical Trial Efficacy Endpoints. Available at https://www.biooncology.com/clinical-trials/efficacy-endpoints.html (Retrieved 2020/08/20), 2016.

[34] E D Yorke, Z Fuks, L Norton, W Whitmore, and C C Ling. Modeling the development of metastases from primary and locally recurrent tumors: comparison with a clinical data base for prostatic cancer. *Cancer research*, 53(13):2987–93, 7 1993.

[35] M. Kimmel and O. Gorlova. Stochastic models of progression of cancer and their use in controlling cancer-related mortality. In *Proceedings of the 2002 American Control Conference (IEEE Cat. No.CH37301)*, pages 3443–3448. IEEE, 2002.

[36] Yen Lin Chia, Peter Salzman, Sylvia K Plevritis, and Peter W Glynn. Simulation-based parameter estimation for complex models: a breast cancer natural history modelling illustration. *Statistical Methods in Medical Research*, 13(6):507–524, 12 2004.

[37] Anne Talkington and Rick Durrett. Estimating Tumor Growth Rates In Vivo. *Bulletin of Mathematical Biology*, 77(10):1934–1954, 10 2015.

[38] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[39] Tom Fawcett. An introduction to roc analysis. *Pattern recognition letters*, 27(8):861–874, 2006.

[40] David R Cox. mregression models and life tables (with discussion). *Journal of the Royal Statistical Society, Series B*, 34(2), 1972.

[41] Hemant Ishwaran, Udaya B Kogalur, Eugene H Blackstone, Michael S Lauer, et al. Random survival forests. *The annals of applied statistics*, 2(3):841–860, 2008.

[42] David Faraggi and Richard Simon. A neural network model for survival data. *Statistics in medicine*, 14(1):73–82, 1995.

[43] Terry M Therneau. *A Package for Survival Analysis in S*, 2015. R package version 2.38.

[44] H. Ishwaran and U.B. Kogalur. *Random Forests for Survival, Regression, and Classification (RF-SRC)*, 2018. R package version 2.6.1.

[45] Frank E Harrell Jr, Robert M Califf, David B Pryor, Kerry L Lee, Robert A Rosati, et al. Evaluating the yield of medical tests. *Jama*, 247(18):2543–2546, 1982.

[46] Marliese Alexander, Rory Wolfe, David Ball, Matthew Conron, Robert G. Stirling, Benjamin Solomon, Michael MacManus, Ann Officer, Sameer Karnam, Kate Burbury, and et al. Lung cancer prognostic index: a risk score to predict overall survival after the diagnosis of non-small-cell lung cancer. *British Journal of Cancer*, 117(5):744–751, Aug 2017.

[47] Joseph Putila, Scot C. Remick, and Nancy Lan Guo. Combining clinical, pathological, and demographic factors refines prognosis of lung cancer: A population-based study. *PLOS ONE*, 6(2):e17493, Feb 2011.

[48] Jie Lin, Corey A Carter, Katherine A McGlynn, Shelia H Zahm, Joel A Nations, William F Anderson, Craig D Shriver, and Kangmin Zhu. A prognostic model to predict mortality among non-small-cell lung cancer patients in the us military health system. *Journal of Thoracic Oncology*, 10(12):1694–1702, 2015.

[49] François Blanchon, Michel Grivaux, Bernard Asselain, François-Xavier Lebas, Jean-Pierre Orlando, Jacques Piquet, and Mahmoud Zureik. 4-year mortality in patients with non-small-cell lung cancer: development and validation of a prognostic index. *The Lancet Oncology*, 7(10):829–836, 2006.

[50] Sumithra J Mandrekar, Steven E Schild, Shauna L Hillman, Katie L Allen, Randolph S Marks, James A Mailliard, James E Krook, Andrew W Maksymiuk, Kari Chansky, Karen Kelly, et al. A prognostic model for advanced stage nonsmall cell lung cancer: pooled analysis of north central cancer treatment group trials. *Cancer*, 107(4):781–792, 2006.

[51] Tien Hoang, Suzanne E Dahlberg, Alan B Sandler, Julie R Brahmer, Joan H Schiller, and David H Johnson. Prognostic models to predict survival in non-small-cell lung cancer patients treated with first-line paclitaxel and carboplatin with or without bevacizumab. *Journal of Thoracic Oncology*, 7(9):1361–1368, 2012.

[52] Benjamin A Derman, Kathryn F Mileham, Philip D Bonomi, Marta Batus, and Mary J Fidler. Treatment of advanced squamous cell carcinoma of the lung: a review. *Translational Lung Cancer Research*, 4(5):524–532, 2015.

[53] Mark Yarchoan, Alexander Hopkins, and Elizabeth M Jaffee. Tumor mutational burden and response rate to pd-1 inhibition. *New England Journal of Medicine*, 377(25):2500–2501, 2017.

[54] Lavinia Tan, Marliese Alexander, Ann Officer, Michael MacManus, Linda Mileshkin, Ross Jennens, Dishan Herath, Richard de Boer, Stephen B Fox, David Ball, et al. Survival difference according to mutation status in a prospective cohort study of australian patients with metastatic non-small-cell lung carcinoma. *Internal Medicine Journal*, 48(1):37–44, 2018.

[55] Michael D Brundage, Diane Davies, and William J Mackillop. Prognostic factors in non-small cell lung cancer: a decade of progress. *Chest*, 122(3):1037–1057, 2002.

[56] Sean Khozin, Gideon M Blumenthal, and Richard Pazdur. Real-world data for clinical evidence generation in oncology. *JNCI: Journal of the National Cancer Institute*, 109(11):djx187, 2017.

[57] Roger Sun, Elaine Johanna Limkin, Maria Vakalopoulou, Laurent Dercle, Stéphane Champiat, Shan Rong Han, Loïc Verlingue, David Brandao, Andrea Lancia, Samy Ammari, et al. A radiomics approach to assess tumour-infiltrating cd8 cells and response to anti-pd-1 or anti-pd-l1 immunotherapy: an imaging biomarker, retrospective multicohort study. *The Lancet Oncology*, 19(9):1180–1191, 2018.

[58] Stefania Rizzo, Sara Raimondi, Evelyn EC de Jong, Wouter van Elmpt, Francesca De Piano, Francesco Petrella, Vincenzo Bagnardi, Arthur Jochems, Massimo Bellomi, Anne Marie Dingemans, et al. Genomics of non-small cell lung cancer (nsclc): Association between ct-based imaging features and egfr and k-ras mutations in 122 patients-an external validation. *European Journal of Radiology*, 110:148–155, 2019.

[59] Hugo JWL Aerts, Emmanuel Rios Velazquez, Ralph TH Leijenaar, Chintan Parmar, Patrick Grossmann, Sara Carvalho, Johan Bussink, René Monshouwer, Benjamin Haibe-Kains, Derek Rietveld, et al. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. *Nature communications*, 5(1):1–9, 2014.

[60] Ahmed Hosny, Chintan Parmar, Thibaud P Coroller, Patrick Grossmann, Roman Zeleznik, Avnish Kumar, Johan Bussink, Robert J Gillies, Raymond H Mak, and Hugo JWL Aerts. Deep learning for lung cancer prognostication: A retrospective multi-cohort radiomics study. *PLoS medicine*, 15(11):e1002711, 2018.

[61] Mohammadhadi Khorrami, Monica Khunger, Alexia Zagouras, Pradnya Patil, Rajat Thawani, Kaustav Bera, Prabhakar Rajiah, Pingfu Fu, Vamsidhar Velcheti, and Anant Madabhushi. Combination of peri-and intratumoral radiomic features on baseline ct scans predicts response to chemotherapy in lung adenocarcinoma. *Radiology: Artificial Intelligence*, 1(2):180012, 2019.

[62] Yiwen Xu, Ahmed Hosny, Roman Zeleznik, Chintan Parmar, Thibaud Coroller, Idalid Franco, Raymond H Mak, and Hugo JWL Aerts. Deep learning predicts lung cancer treatment response from serial medical imaging. *Clinical Cancer Research*, 25(11):3266–3275, 2019.

[63] Philippe Lambin, Ralph TH Leijenaar, Timo M Deist, Jurgen Peerlings, Evelyn EC De Jong, Janita Van Timmeren, Sebastian Sanduleanu, Ruben THM

Larue, Aniek JG Even, Arthur Jochems, et al. Radiomics: the bridge between medical imaging and personalized medicine. *Nature reviews Clinical oncology*, 14(12):749–762, 2017.

[64] Keehyun Earm and Yung E Earm. Integrative approach in the era of failing drug discovery and development. *Integrative medicine research*, 3(4):211–216, 2014.

[65] Informa. Pharmaceutical clinical trial intelligence products. Available at https://pharmaintelligence.informa.com/products-and-services/data-and-analysis/citeline-joins-informas-pharma-intelligence (Retrieved 2016-12-05), 2016.

[66] Laeeq Malik, Alex Mejia, Helen Parsons, Benjamin Ehler, Devalingam Mahalingam, Andrew Brenner, John Sarantopoulos, and Steven Weitman. Predicting success in regulatory approval from phase i results. *Cancer chemotherapy and pharmacology*, 74(5):1099–1103, 2014.

[67] John Goffin, Stefan Baral, Dongsheng Tu, Dora Nomikos, and Lesley Seymour. Objective responses in patients with malignant melanoma or renal cell cancer in early clinical studies do not predict regulatory approval. *Clinical cancer research*, 11(16):5928–5934, 2005.

[68] Robert H El-Maraghi and Elizabeth A Eisenhauer. Review of phase ii trial designs used in studies of molecular targeted agents: outcomes and predictors of success in phase iii. *Journal of Clinical Oncology*, 26(8):1346–1354, 2008.

[69] Denis L Jardim, Eric S Groves, Philip P Breitfeld, and Razelle Kurzrock. Factors associated with failure of oncology drugs in late-stage clinical development: a systematic review. *Cancer treatment reviews*, 52:12–21, 2017.

[70] JA DiMasi, JC Hermann, K Twyman, RK Kondru, S Stergiopoulos, KA Getz, and W Rackoff. A tool for predicting regulatory approval after phase ii testing of new oncology compounds. *Clinical Pharmacology & Therapeutics*, 98(5):506–513, 2015.

[71] Deborah A Zarin, Tony Tse, Rebecca J Williams, and Sarah Carr. Trial reporting in clinicaltrials. gov—the final rule. *New England Journal of Medicine*, 375(20):1998–2004, 2016.

[72] Donald B Rubin. Inference and missing data. *Biometrika*, pages 581–592, 1976.

[73] Stef Van Buuren. *Flexible imputation of missing data*. CRC press, 2012.

[74] Roderick JA Little and Donald B Rubin. *Statistical analysis with missing data*. John Wiley & Sons, 2014.

[75] Joseph L Schafer. *Analysis of incomplete multivariate data*. CRC press, 1997.

[76] Donald B Rubin. *Multiple imputation for nonresponse in surveys*, volume 81. John Wiley & Sons, 2004.

[77] Craig K Enders. *Applied missing data analysis*. Guilford Press, 2010.

[78] Ross Quinlan. C5.0: An informal tutorial, 1998.

[79] Linda M Collins, Joseph L Schafer, and Chi-Ming Kam. A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological methods*, 6(4):330, 2001.

[80] Donald B Rubin. Multiple imputation after 18+ years. *Journal of the American statistical Association*, 91(434):473–489, 1996.

[81] Joseph L Schafer and John W Graham. Missing data: our view of the state of the art. *Psychological methods*, 7(2):147, 2002.

[82] M Kuhn, S Weston, N Coulter, and R Quinlan. *C50: C5.0 decision trees and rule-based models*, 2014. R package version 0.1.0-21.

[83] Leo Breiman, Jerome Friedman, Charles J Stone, and Richard A Olshen. *Classification and regression trees*. CRC press, 1984.

[84] Angela K Green, Katherine E Reeder-Hayes, Robert W Corty, Ethan Basch, Mathew I Milowsky, Stacie B Dusetzina, Antonia V Bennett, and William A Wood. The project data sphere initiative: accelerating cancer research by sharing data. *The oncologist*, 20(5):464, 2015.

[85] Barbara E Bierer, Rebecca Li, Mark Barnes, and Ida Sim. A global, neutral platform for sharing trial data. *New England Journal of Medicine*, 374(25):2411–2413, 2016.

[86] Jillian Deutsch. Von der leyen: Life won't return to normal until vaccine. *Politico*, Apr 2020. Accessed: 2020-06-25.

[87] Yasmeen Abutaleb, Josh Dawsey, Laurie McGinley, and Carolyn Y Johnson. Trump pushing officials to speed up already-ambitious coronavirus vaccine timeline. *Washington Post*, Jun 2020. Accessed: 2020-06-15.

[88] Nicole Lurie, Melanie Saville, Richard Hatchett, and Jane Halton. Developing covid-19 vaccines at pandemic speed. *New England Journal of Medicine*, 382(21):1969–1973, 2020.

[89] Thomas M. Burton. Fda to require proof virus vaccine is effective before approving its use. *The Wall Street Journal*, Jun 2020.

[90] International Federation of Pharmaceutical Manufacturers & Associations. The complex journey of a vaccine. Technical report, International Federation of Pharmaceutical Manufacturers & Associations, Jul 2019.

[91] Jozef Nauta. *Statistics in Clinical and Observational Vaccine Studies*. Springer, 2020.

[92] U.S. Food and Drug Administration. Development and licensure of vaccines to prevent covid-19. Available at https://www.fda.gov/media/139638/download (Retrieved 2020-06-29), 2020.

[93] Amy Caryn Sherman, Aneesh Mehta, Neal W Dickert, Evan J Anderson, and Nadine Rouphael. The future of flu: a review of the human challenge model and systems biology for advancement of influenza vaccinology. *Frontiers in cellular and infection microbiology*, 9:107, 2019.

[94] Debbie-Ann T Shirley and Monica A McArthur. The utility of human challenge studies in vaccine development: lessons learned from cholera. *Vaccine: development and therapy*, 2011(1):3, 2011.

[95] Danielle I Stanisic, James S McCarthy, and Michael F Good. Controlled human malaria infection: applications, advances, and challenges. *Infection and immunity*, 86(1), 2018.

[96] Meriel Raymond, Malick M Gibani, Nicholas PJ Day, and Phaik Yeong Cheah. Typhoidal salmonella human challenge studies: ethical and practical challenges and considerations for low-resource settings. *Trials*, 20(2):1–7, 2019.

[97] Wudan Yan. *Challenge accepted: Human challenge trials for dengue*. Nature Publishing Group, 2015.

[98] Seema K Shah, Franklin G Miller, Thomas C Darton, Devan Duenas, Claudia Emerson, Holly Fernandez Lynch, Euzebiusz Jamrozik, Nancy S Jecker, Dorcas Kamuya, Melissa Kapulu, et al. Ethics of controlled human infection to address covid-19. *Science*, 368(6493):832–834, 2020.

[99] Moderna. Moderna advances late-stage development of its vaccine (mrna-1273) against covid-19. Available at https://investors.modernatx.com/news-releases/news-release-details/moderna-advances-late-stage-development-its-vaccine-mrna-1273 (Retrieved 2020-06-29), 2020.

[100] AstraZeneca. Astrazeneca advances response to global covid-19 challenge as it receives first commitments for oxford's potential new vaccine. Available at https://www.astrazeneca.com/media-centre/press-releases/2020/astrazeneca-advances-response-to-global-covid-19-challenge-as-it-receives-first-commitments-for-oxfords-potential-new-vaccine.html (Retrieved 2020-06-29), 2020.

[101] Business Wire. Pfizer and biontech granted fda fast track designation for two investigational mrna-based vaccine candidates against sars-cov-2. Available at https://www.businesswire.com/news/home/20200713005168/en/Pfizer-BioNTech-Granted-FDA-Fast-Track-Designation (Retrieved 2020-06-29), 2020.

[102] Moderna. Moderna announces ind submitted to u.s. fda for phase 2 study of mrna vaccine (mrna-1273) against novel coronavirus. Available at https://investors.modernatx.com/news-releases/news-release-details/moderna-announces-ind-submitted-us-fda-phase-2-study-mrna (Retrieved 2020-06-29), 2020.

[103] Nick Paul Taylor. Astrazeneca's covid-19 vaccine enters phase 2/3 clinical trial. Available at https://www.fiercebiotech.com/biotech/astrazeneca-s-covid-19-vaccine-enters-phase-2-3-clinical-trial (Retrieved 2020-06-29), 2020.

[104] Christopher Jennison and Bruce W Turnbull. Group-sequential analysis incorporating covariate information. *Journal of the American Statistical Association*, 92(440):1330–1341, 1997.

[105] Moderna. Cove study: Participate to make a world of difference. Available at https://www.modernatx.com/cove-study (Retrieved 2020-11-20), 2020.

[106] Centers for Disease Control and Prevention. Covid-19 provisional counts - weekly updates by select demographic and geographic characteristics. Available at https://www.cdc.gov/nchs/nvss/vsrr/covid_weekly/index.htm (Retrieved 2020-06-29), 2020.

[107] Graziano Onder, Giovanni Rezza, and Silvio Brusaferro. Case-fatality rate and characteristics of patients dying in relation to covid-19 in italy. *Jama*, 323(18):1775–1776, 2020.

[108] World Health Organization. Feasibility, potential value and limitations of establishing a closely monitored challenge model of experimental covid-19 infection and illness in healthy young adult volunteers. Technical report, World Health Organization, Jun 2020.

[109] Stephen A Lauer, Kyra H Grantz, Qifang Bi, Forrest K Jones, Qulu Zheng, Hannah R Meredith, Andrew S Azman, Nicholas G Reich, and Justin Lessler. The incubation period of coronavirus disease 2019 (covid-19) from publicly reported confirmed cases: estimation and application. *Annals of internal medicine*, 172(9):577–582, 2020.

[110] Wei Jie Guan, Zheng Yi Ni, Yu Hu, Wen Hua Liang, Chun Quan Ou, Jian Xing He, Lei Liu, Hong Shan, Chun Liang Lei, David SC Hui, et al. Clinical characteristics of coronavirus disease 2019 in china. *New England journal of medicine*, 382(18):1708–1720, 2020.

[111] Qun Li, Xuhua Guan, Peng Wu, Xiaoye Wang, Lei Zhou, Yeqing Tong, Ruiqi Ren, Kathy SM Leung, Eric HY Lau, Jessica Y Wong, et al. Early transmission dynamics in wuhan, china, of novel coronavirus–infected pneumonia. *New England Journal of Medicine*, 2020.

[112] Robert D Kirkcaldy, Brian A King, and John T Brooks. Covid-19 and postinfection immunity: Limited evidence, many remaining questions. *JAMA*, 2020.

[113] Jeffrey Seow, Carl Graham, Blair Merrick, Sam Acors, Kathryn J.A. Steel, Oliver Hemmings, Aoife O'Bryne, Neophytos Kouphou, Suzanne Pickering, Rui Galao, Gilberto Betancor, Harry D Wilson, Adrian W Signell, Helena Winstone, Claire Kerridge, Nigel Temperton, Luke Snell, Karen Bisnauthsing, Amelia Moore, Adrian Green, Lauren Martinez, Brielle Stokes, Johanna Honey, Alba Izquierdo-Barras, Gill Arbane, Amita Patel, Lorcan OConnell, Geraldine O Hara, Eithne MacMahon, Sam Douthwaite, Gaia Nebbia, Rahul Batra, Rocio Martinez-Nunez, Jonathan D. Edgeworth, Stuart J.D. Neil, Michael H. Malim, and Katie Doores. Longitudinal evaluation and decline of antibody responses in sars-cov-2 infection. Available at https://www.medrxiv.org/content/10.1101/2020.07.09.20148429v1 (Retrieved 2020-06-29), 2020.

[114] Conor P Farrington and Godfrey Manning. Test statistics and sample size formulae for comparative binomial trials with null hypothesis of non-zero risk difference or non-unity relative risk. *Statistics in medicine*, 9(12):1447–1454, 1990.

[115] Joseph L Fleiss, Bruce Levin, and Myunghee Cho Paik. *Statistical methods for rates and proportions*. John Wiley & Sons, 2013.

[116] Stuart J Pocock. Group sequential methods in the design and analysis of clinical trials. *Biometrika*, 64(2):191–199, 1977.

[117] Jesus Fernandez-Villaverde and Charles I Jones. Estimating and simulating a sird model of covid-19 for many countries, states, and cities. Technical report, National Bureau of Economic Research, 2020.

[118] Center for Systems Science and Engineering (CSSE). COVID-19 Data Repository by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University. Available at https://github.com/CSSEGISandData/COVID-19 (Retrieved 2020-06-29), 2020.

[119] Ensheng Dong, Hongru Du, and Lauren Gardner. An interactive web-based dashboard to track covid-19 in real time. *The Lancet infectious diseases*, 20(5):533–534, 2020.

[120] John Cochrane. An sir model with behavior. Available at https://johnhcochrane.blogspot.com/2020/05/an-sir-model-with-behavior.html (Retrieved 2020-06-29), 2020.

[121] Moderna. Moderna and catalent announce collaboration for fill-finish manufacturing of moderna's covid-19 vaccine candidate. Available at https://investors.modernatx.com/news-releases/news-release-details/moderna-and-catalent-announce-collaboration-fill-finish (Retrieved 2020-06-29), 2020.

[122] Emergent BioSolutions. Emergent biosolutions signs agreement to be u.s. manufacturing partner for astrazeneca's covid-19 vaccine candidate. Available at https://investors.emergentbiosolutions.com/news-releases/news-release-details/emergent-biosolutions-signs-agreement-be-us-manufacturing-0 (Retrieved 2020-06-29), 2020.

[123] Vahid Montazerhodjat, Shomesh E Chaudhuri, Daniel J Sargent, and Andrew W Lo. Use of bayesian decision analysis to minimize harm in patient-centered randomized clinical trials in oncology. *JAMA oncology*, 3(9):e170123–e170123, 2017.

[124] Leah Isakov, Andrew W Lo, and Vahid Montazerhodjat. Is the fda too conservative or too aggressive?: A bayesian decision analysis of clinical trial design. *Journal of econometrics*, 211(1):117–136, 2019.

[125] Shomesh E Chaudhuri, Martin P Ho, Telba Irony, Murray Sheldon, and Andrew W Lo. Patient-centered clinical trials. *Drug discovery today*, 23(2):395–401, 2018.

[126] Anthony T Newall, Nathorn Chaiyakunapruk, Philipp Lambach, and Raymond CW Hutubessy. Who guide on the economic evaluation of influenza vaccination. *Influenza and other respiratory viruses*, 12(2):211–219, 2018.

[127] Lisa A Jackson, Evan J Anderson, Nadine G Rouphael, Paul C Roberts, Mamodikoe Makhene, Rhea N Coler, Michele P McCullough, James D Chappell, Mark R Denison, Laura J Stevens, et al. An mrna vaccine against sars-cov-2—preliminary report. *New England Journal of Medicine*, 383(20):1920–1931, 2020.

[128] Pedro M Folegatti, Katie J Ewer, Parvinder K Aley, Brian Angus, Stephan Becker, Sandra Belij-Rammerstorfer, Duncan Bellamy, Sagida Bibi, Mustapha Bittaye, Elizabeth A Clutterbuck, et al. Safety and immunogenicity of the chadox1 ncov-19 vaccine against sars-cov-2: a preliminary report of a phase 1/2, single-blind, randomised controlled trial. *The Lancet*, 396(10249):467–478, 2020.

[129] Ben Bambery, Michael Selgelid, Charles Weijer, Julian Savulescu, and Andrew J Pollard. Ethical criteria for human challenge studies in infectious diseases. *Public Health Ethics*, 9(1):92–103, 2016.

[130] G Miller Franklin and Christine Grady. The ethical challenge of infection-inducing challenge experiments. *Clinical Infectious Diseases*, 33(7):1028–1033, 2001.

[131] Nir Eyal, Marc Lipsitch, and Peter G Smith. Human challenge studies to accelerate coronavirus vaccine licensure. *The Journal of infectious diseases*, 221(11):1752–1756, 2020.

[132] Stanley A Plotkin and Arthur Caplan. Extraordinary diseases require extraordinary solutions. *Vaccine*, 38(24):3987, 2020.

[133] Euzebiusz Jamrozik and Michael J Selgelid. Covid-19 human challenge studies: ethical issues. *The Lancet Infectious Diseases*, 2020.

[134] World Health Organization. Key criteria for the ethical acceptability of covid-19 human challenge studies. Technical report, World Health Organization, 2020.

[135] Meagan E Deming, Nelson L Michael, Merlin Robb, Myron S Cohen, and Kathleen M Neuzil. Accelerating development of sars-cov-2 vaccines—the role for controlled human infection models. *New England Journal of Medicine*, 2020.

[136] Jeremy Bentham. *A fragment on government*. The Lawbook Exchange, Ltd., 2001.

[137] 1Day Sooner. Covid-19 human challenge trials. Available at https://1daysooner.org/ (Retrieved 2020-07-27), 2020.

[138] Owermohle Sarah. White house pressure for a vaccine raises risk the u.s. will approve one that doesn't work. Available at https://www.politico.com/news/2020/06/15/pressure-coronavirus-vaccine-risk-approval-316094 (Retrieved 2020-06-29), 2020.

[139] Gupta, Sanjay and Cohen, Elizabeth and Howard, Jacqueline. Us considering coronavirus strain for potentialhuman challenge trials. Available at https://www.cnn.com/2020/08/14/health/coronavirus-vaccine-strain-us-scientists-bn/index.html (Retrieved 2020-11-20), 2020.

[140] Steenhuysen, Julie. U.s. to make coronavirus strain for possible human challenge trials. Available at https://www.reuters.com/article/us-health-coronavirus-vaccine-challenge/exclusive-u-s-to-make-coronavirus-strain-for-possible-human-challenge-trials-idUSKCN25A1EL (Retrieved 2020-11-20), 2020.

[141] World Health Organization. An international randomised trial of candidate vaccines against covid-19. Technical report, World Health Organization, April 2020.

[142] Alex Keown. "operation warp speed" narrows list of potential covid-19 vaccine candidates down to five. Available at https://www.biospace.com/article/white-house-narrows-covid-19-vaccine-candidates-down-to-5-for-operation-warp-speed/ (Retrieved 2020-06-29), 2020.

[143] Jared S Hopkins and Peter Loftus. Coronavirus researchers compete to enroll subjects for vaccine tests. Available at https://www.wsj.com/articles/coronavirus-researchers-compete-to-enroll-subjects-for-vaccine-tests-11593968711 (Retrieved 2020-06-29), 2020.

[144] Pfizer. Two cities in brazil to host late-stage clinical trials for pfizer and biontech's mrna vaccine against covid-19. Available at https://www.pfizer.com.co/tu-salud/two-cities-brazil-host-late-stage-clinical-trials-pfizer-and-biontechs-mrna-vaccine-against-covid (Retrieved 2020-11-20), 2020.

[145] Debashree Ray, Maxwell Salvatore, Rupam Bhattacharyya, Lili Wang, Jiacong Du, Shariq Mohammed, Soumik Purkayastha, Aritra Halder, Alexander Rix, Daniel Barker, Michael Kleinsasser, Yiwang Zhou, Debraj Bose, Peter Song, Mousumi Banerjee, Veerabhadran Baladandayuthapani, Parikshit Ghosh, and Bhramar Mukherjee. Predictions, role of interventions and effects of a historic national lockdown in india's response to the the covid-19 pandemic: Data science call to arms. *Harvard Data Science Review*, 6 2020. https://hdsr.mitpress.mit.edu/pub/r1qq01kw.

[146] Bernard Munos. Lessons from 60 years of pharmaceutical innovation. *Nature Reviews Drug discovery*, 8(12):959–968, 2009.

[147] David C Mowery, Richard R Nelson, Bhaven N Sampat, and Arvids A Ziedonis. The growth of patenting and licensing by us universities: an assessment of the effects of the bayh–dole act of 1980. *Research policy*, 30(1):99–119, 2001.

[148] Association of University Technology Managers. Autm u.s. licensing activity survey fy2016. Technical report, 2016.

[149] Eric G Campbell, Joshua B Powers, David Blumenthal, and Brian Biles. Inside the triple helix: Technology transfer and commercialization in the life sciences. *Health Affairs*, 23(1):64–76, 2004.

[150] David Mowery, Richard Nelson, Bhavan Sampat, and Arvids Ziedonis. The effects of the bayh-dole act on us university research and technology transfer. *Industrializing knowledge*, pages 269–306, 1999.

[151] Asher Mullard. 2016 fda drug approvals. *Nature Reviews Drug discovery*, 16:73–76, 2017.

[152] Fredric J Cohen. Macro trends in pharmaceutical innovation. *Nature Reviews Drug Discovery*, 4(1):78–84, 2005.

[153] Michael Lanthier, Kathleen L Miller, Clark Nardinelli, and Janet Woodcock. An improved approach to measuring drug innovation finds steady rates of first-in-class pharmaceuticals, 1987–2011. *Health Affairs*, 32(8):1433–1439, 2013.

[154] U.S. Food and Drug Administration. Approved drug products with therapeutic equivalence evaluations (orange book). Available at http://www.fda.gov/drugs/drug-approvals-and-databases/approved-drug-products-therapeutic-equivalence-evaluations-orange-book (Retrieved 2020/08/20), 2020.

[155] Ashley J Stevens, Jonathan J Jensen, Katrine Wyller, Patrick C Kilgore, Sabarni Chatterjee, and Mark L Rohrbaugh. The role of public-sector research in the discovery of drugs and vaccines. *New England Journal of Medicine*, 364(6):535–541, 2011.

[156] Pablo Legorreta. Royalty pharma: Transforming the funding of life sciences through collaborative capital. Presentation at MIT 15.480 "The Science and Business of Biotechnology", 2019.

[157] Charles Weissmann. Recombinant interferon-the 20th anniversary. In *Recombinant protein drugs*, pages 3–41. Springer, 2001.

[158] Andrew Pollack. Eugene goldwasser, biochemist behind an anemia drug, dies at 88. Available at https://www.nytimes.com/2010/12/21/health/21goldwasser.html (Retrieved 2020/08/20), 2010.

[159] The Associated Press. Pfizer settles b.y.u. lawsuit over development of celebrex. Available at https://www.nytimes.com/2012/05/02/health/pfizer-settles-byu-lawsuit-over-development-of-celebrex.html (Retrieved 2020/08/20), 2012.

[160] EvaluatePharma. World preview 2016, outlook to 2022. Available at https://info.evaluategroup.com/rs/607-YGS-364/images/wp16.pdf (Retrieved 2020/08/20), 2016.

[161] Weizmann Institute of Science. Cop 1 (copaxone): The story of a drug. Available at https://wis-wander.weizmann.ac.il/life-sciences/cop-1-copaxone%C2%AE-story-drug (Retrieved 2020/08/20), 1997.

[162] Rebecca Robbins. The lab breakthrough that paved the way for hepatitis c cures. Available at https://www.statnews.com/2016/09/13/lab-breakthrough-hepatitis-c/ (Retrieved 2020/08/20), 2016.

[163] Denise Heady. Ucla's dennis slamon awarded sjoberg prize for pioneering cancer research. Available at https://newsroom.ucla.edu/releases/dennis-slamon-sjoberg-prize-pioneering-cancer-research (Retrieved 2020/08/20), 2019.

[164] Heidi Williams. Orange book patent and exclusivity data. Available at https://data.nber.org/fda/orange-book/historical/1986-2016/ (Retrieved 2020/08/20), 2019.

[165] Jean Roth. Fda orange book. Available at https://data.nber.org/data/fda-orange-book-data.html (Retrieved 2020/08/20), 2018.

[166] United States Patent and Trademark Office. Patent full-text database. Available at http://patft.uspto.gov/netahtml/PTO/index.html (Retrieved 2020/08/20).

[167] United States Patent and Trademark Office. Patent assignment search. Available at https://assignment.uspto.gov/patent/index.html#/patent/search (Retrieved 2020/08/20).

[168] U.S. Food and Drug Administration. Indevus pharmaceuticals, inc.; withdrawal of approval of a new drug application. Available at https://www.federalregister.gov/documents/2007/01/30/E7-1414/indevus-pharmaceuticals-inc-withdrawal-of-approval-of-a-new-drug-application (Retrieved 2020/08/20), 2007.

[169] Steven Prokesch. The edison of medicine. Available at https://hbr.org/2017/03/the-edison-of-medicine (Retrieved 2020/08/20), 2017.

[170] Matthew Herper. Lilly gives away prozac, makes money. Available at https://www.forbes.com/2001/11/16/1116lilly.html#94c3aef35ee3 (Retrieved 2020/08/20), 2001.

[171] Chris Morrison and Riku Lähteenmäki. Public biotech in 2014-the numbers. *Nature Biotechnology*, 33(7):703–709, 2015.

[172] Merck. Merck to acquire smartcells, inc. Available at https://www.mrknewsroom.com/press-release/corporate-news/merck-acquire-smartcells-inc (Retrieved 2020/08/20), 2010.

[173] Rob Matheson. The half-billion-dollar idea. Available at http://news.mit.edu/2013/todd-zion-smartcells-0408 (Retrieved 2020/08/20), 2013.

[174] Ben Fidler. Alkermes spinoff civitas, poised for ipo, sells to acorda for $525m. Available at https://xconomy.com/boston/2014/09/24/alkermes-spinoff-civitas-poised-for-ipo-sells-to-acorda-for-525m/ (Retrieved 2020/08/20), 2014.

[175] U.S. Food and Drug Administration. 2017 new drug therapy approvals. Available at https://www.fda.gov/downloads/AboutFDA/CentersOffices/OfficeofMedicalProductsandTobacco/CDER/ReportsBudgets/UCM591976.pdf (Retrieved 2020/08/20), 2018.

[176] U.S. Food and Drug Administration. Priority review. Available at http://www.fda.gov/patients/fast-track-breakthrough-therapy-accelerated-approval-priority-review/priority-review (Retrieved 2020/08/20), 2018.

[177] Robert Weisman. Takeda gets fda approval for multiple myeloma pill. Available at https://www.bostonglobe.com/business/2015/11/20/takeda-gets-fda-approval-for-multiple-myeloma-pill-developed-cambridge/YRxwHUrcljJ3pY5iS6VvtJ/story.html (Retrieved 2020/08/20), 2015.

[178] Frederick J Raal, Robert P Giugliano, Marc S Sabatine, Michael J Koren, Dirk Blom, Nabil G Seidah, Narimon Honarpour, Armando Lira, Allen Xue, Padmaja Chiruvolu, et al. Pcsk9 inhibition-mediated reduction in lp(a) with evolocumab: an analysis of 10 clinical trials and the ldl receptor's role. *Journal of lipid research*, 57(6):1086–1096, 2016.

[179] Alnylam Pharmaceuticals. Alnylam announces first-ever fda approval of an rnai therapeutic, onpattrotm (patisiran) for the treatment of the polyneuropathy of hereditary transthyretin-mediated amyloidosis in adults. Available at https://investors.alnylam.com/press-release?id=22946 (Retrieved 2020/08/20), 2018.

[180] Francis P Tally and Michael F DeBruin. Development of daptomycin for gram-positive infections. *Journal of Antimicrobial Chemotherapy*, 46(4):523–526, 2000.

[181] Matthew Herper. Biotech battle scars. Available at https://www.forbes.com/2007/09/14/biotechnology-venture-capital-biz-sci-cx_mh_0914venture.html#20de6b2b681a (Retrieved 2020/08/20), 2007.

[182] Rochelle Tung. Mit related to top biotech companies. Available at http://tech.mit.edu/V115/N67/biotech.67n.html (Retrieved 2020/08/20), 1996.

[183] MIT News Office. Study traces impact of government-funded mit research on biotechnology and drugs. Available at http://news.mit.edu/1996/biotechmitgov (Retrieved 2020/08/20), 1996.

[184] Verastem. Verastem secures $16 million in series a financing for novel cancer stem cell drugs. Available at https://www.businesswire.com/news/home/20101116005654/en/Verastem-Secures-16-Million-Series-Financing-Cancer (Retrieved 2020/08/20), 2010.

[185] Alnylam Pharmaceuticals. Alnylam announces fda acceptance of new drug application (nda) and priority review status for patisiran, an investigational rnai therapeutic for the treatment of hereditary attr (hattr) amyloidosis. Available at https://www.businesswire.com/news/home/20180201006407/en/Alnylam-Announces-FDA-Acceptance-New-Drug-Application (Retrieved 2020/08/20), 2018.

[186] Duke University. Duke university & deerfield management announce four points innovation. Available at https://olv.duke.edu/news/duke-deerfield-announce-four-points-innovation/ (Retrieved 2020/08/20), 2019.

[187] MPEG LA. Mpeg la initiative to address crispr licensing. Available at https://www.businesswire.com/news/home/20161206006182/en/MPEG-LA-Initiative-Address-CRISPR-Licensing (Retrieved 2020/08/20), 2016.

[188] Stacy Lawrence. Celgene backs four nci cancer centers with $50m to create research consortium. Available at https://www.fiercebiotech.com/biotech/celgene-backs-four-nci-cancer-centers-50m-to-create-research-consortium (Retrieved 2020/08/20), 2016.

[189] David E Fagnan, Austin A Gromatzky, Roger M Stein, Jose-Maria Fernandez, and Andrew W Lo. Financing drug discovery for orphan diseases. *Drug discovery today*, 19(5):533–538, 2014.

[190] David E Fagnan, N Nora Yang, John C McKew, and Andrew W Lo. Financing translation: Analysis of the ncats rare-diseases portfolio. *Science translational medicine*, 7(276):276ps3–276ps3, 2015.

[191] Paul D Maher. Pharmacogenomics of rare and monogenic disorders. In *Pharmacogenomics and Personalized Medicine*, pages 479–497. Springer, 2008.

[192] Svenja Hager. Explaining the implied correlation smile. *Pricing Portfolio Credit Derivatives by Means of Evolutionary Algorithms*, pages 41–71, 2008.

[193] John C Hull and Alan D White. Valuation of a cdo and an n-th to default cds without monte carlo simulation. *The Journal of Derivatives*, 12(2):8–23, 2004.

[194] Oldrich Alfons Vasicek. *Probability of loss on loan portfolio*. KMV, 1987.

[195] David X Li. On default correlation: A copula function approach. *The Journal of Fixed Income*, 9(4):43–54, 2000.

[196] David E Fagnan, Jose Maria Fernandez, Andrew W Lo, and Roger M Stein. Can financial engineering cure cancer? *American Economic Review*, 103(3):406–11, 2013.

[197] Vahid Montazerhodjat, John J Frishkopf, and Andrew W Lo. Financing drug discovery via dynamic leverage. *Drug discovery today*, 21(3):410–414, 2016.

[198] GW Pharmaceuticals plc. GW Pharmaceuticals plc Announces the Sale of Priority Review Voucher for $105M. Available at https://www.globenewswire.com/news-release/2019/03/18/1756217/0/en/GW-Pharmaceuticals-plc-Announces-the-Sale-of-Priority-Review-Voucher-for-105M.html (Retrieved 2020/08/20), 2019.

[199] Carlos E. Ortiz, Charles A. Stone, and Anne Zissu. Pattern Risk of the Securitized Biopharmaceutical Mega-Fund. *The Journal of Structured Finance*, page jsf.2020.1.103, jun 2020.

[200] Andrew W Lo, Carole Ho, Jayna Cummings, and Kenneth S Kosik. Parallel discovery of alzheimer's therapeutics. *Science translational medicine*, 6(241):241cm5, 2014.

[201] Sonya Das, Raphaël Rousseau, Peter C Adamson, and Andrew W Lo. New business models to accelerate innovation in pediatric oncology therapeutics: a review. *JAMA oncology*, 4(9):1274–1280, 2018.

[202] Shomesh Chaudhuri, Katherine Cheng, Andrew W Lo, Shirley Pepke, Sergio Rinaudo, Lynda Roman, and Ryan Spencer. A portfolio approach to accelerate therapeutic innovation in ovarian cancer. *Journal of Investment Management*, 17(2):5–16, 2019.

[203] Global Coalition for Adaptive Research. A trial to evaluate multiple regimens in newly diagnosed and recurrent glioblastoma (gbm agile). Available at https://clinicaltrials.gov/ct2/show/NCT03970447 (Retrieved 2020/01/13), 2020.

[204] Quinn T Ostrom, Haley Gittleman, Peter Liao, Toni Vecchione-Koval, Yingli Wolinsky, Carol Kruchko, and Jill S Barnholtz-Sloan. Cbtrus statistical report: primary brain and other central nervous system tumors diagnosed in the united states in 2010–2014. *Neuro-oncology*, 19(suppl_5):v1–v88, 2017.

[205] Mary Elizabeth Davis. Glioblastoma: overview of disease and treatment. *Clinical journal of oncology nursing*, 20(5):S2, 2016.

[206] Jigisha P Thakkar, Therese A Dolecek, Craig Horbinski, Quinn T Ostrom, Donita D Lightner, Jill S Barnholtz-Sloan, and John L Villano. Epidemiologic and molecular prognostic review of glioblastoma. *Cancer Epidemiology and Prevention Biomarkers*, 23(10):1985–1996, 2014.

[207] Donald A Berry. Adaptive clinical trials in oncology. *Nature reviews Clinical oncology*, 9(4):199, 2012.

[208] Chi Heem Wong, Kien Wei Siah, and Andrew W Lo. Estimating clinical trial success rates and related parameters in oncology. *Available at SSRN 3355022*, 2019.

[209] Steven M Paul, Daniel S Mytelka, Christopher T Dunwiddie, Charles C Persinger, Bernard H Munos, Stacy R Lindborg, and Aaron L Schacht. How to improve r&d productivity: the pharmaceutical industry's grand challenge. *Nature Reviews Drug Discovery*, 9(3):203–214, 2010.

[210] Jeffrey Strovel, Sitta Sittampalam, Nathan P Coussens, Michael Hughes, James Inglese, Andrew Kurtz, Ali Andalibi, Lavonne Patton, Chris Austin, Michael Baltezor, et al. Early drug discovery and development guidelines: for academic researchers, collaborators, and start-up companies. In *Assay Guidance Manual [Internet]*. Eli Lilly & Company and the National Center for Advancing Translational Sciences, 2016.

[211] Kenneth I Kaitin and Joseph A DiMasi. Pharmaceutical innovation in the 21st century: new drug approvals in the first decade, 2000–2009. *Clinical pharmacology & therapeutics*, 89(2):183–188, 2011.

[212] Janet Woodcock and Lisa M LaVange. Master protocols to study multiple therapies, multiple diseases, or both. *New England Journal of Medicine*, 377(1):62–70, 2017.

[213] Houduo Qi and Defeng Sun. A quadratically convergent newton method for computing the nearest correlation matrix. *SIAM journal on matrix analysis and applications*, 28(2):360–385, 2006.

[214] BCC Research. Brain tumor therapeutics: Global markets to 2023. Available at https://www.bccresearch.com/market-research/pharmaceuticals/brain-tumor-therapeutics-markets.html (Retrieved 2020/08/25), 2019.

[215] America's Health Insurance Plans Center for Policy and Research. High-priced drugs: Estimates of annual per-patient expenditures for 150 specialty medications. Available at https://www.ahip.org/wp-content/uploads/2016/04/HighPriceDrugsReport.pdf (Retrieved 2020/08/25), 2016.

[216] Lai-San Tham, Lingzhi Wang, Ross A Soo, Soo-Chin Lee, How-Sung Lee, Wei-Peng Yong, Boon-Cher Goh, and Nicholas H G Holford. A Pharmacodynamic Model for the Time Course of Tumor Shrinkage by Gemcitabine + Carboplatin in Non-Small Cell Lung Cancer Patients. *Clinical Cancer Research*, 14(13):4213–8, 7 2008.

[217] DR Mould, A-C Walz, T Lave, JP Gibbs, and B Frame. Developing Exposure/Response Models for Anticancer Drug Treatment: Special Considerations. *CPT: Pharmacometrics & Systems Pharmacology*, 4(1):12–27, 1 2015.

[218] Stef Vab Buuren and Karin Groothuis-Oudshoorn. mice: Multivariate imputation by chained equations in r. *Journal of statistical software*, 45(3), 2011.

[219] Donald W Light, Jon Kim Andrus, and Rebecca N Warburton. Estimated research and development costs of rotavirus vaccines. *Vaccine*, 27(47):6627–6633, 2009.

[220] Arianna Waye, Philip Jacobs, and Anthony B Schryvers. Vaccine development costs: a review. *Expert review of vaccines*, 12(12):1495–1501, 2013.

[221] Thomas J Moore, Hanzhe Zhang, Gerard Anderson, and G Caleb Alexander. Estimated costs of pivotal trials for novel therapeutic agents approved by the us food and drug administration, 2015-2016. *JAMA internal medicine*, 178(11):1451–1457, 2018.

[222] U.S. Food and Drug Administration. Orange book preface. Available at http://www.fda.gov/drugs/development-approval-process-drugs/orange-book-preface (Retrieved 2020/08/20), 2020.

[223] FierceBiotech. Conatus pharma aims for $69m ipo to fund liver-disease drug. Available at https://www.fiercebiotech.com/venture-capital/conatus-

pharma-aims-for-69m-ipo-to-fund-liver-disease-drug (Retrieved 2020/08/20), 2013.

[224] U.S. Food and Drug Administration. Nda and bla approval times. Available at http://www.fda.gov/drugs/nda-and-bla-approvals/nda-and-bla-approval-times (Retrieved 2020/08/20), 2017.

[225] U.S. Food and Drug Administration. Nda and bla calendar year approvals. Available at http://www.fda.gov/drugs/nda-and-bla-approvals/nda-and-bla-calendar-year-approvals (Retrieved 2020/08/20), 2019.

[226] Nicholas Wade. In new way to edit dna, hope for treating disease. Available at https://www.nytimes.com/2009/12/29/health/research/29zinc.html (Retrieved 2020/08/20), 2009.

[227] Jin-soo Kim and Carl O Pabo. Poly zinc finger proteins with improved linkers, 2002. US Patent 6,479,626.