Financial and Analytic Innovations for Therapeutic Development

by

Qingyang Xu

B.S., Stanford University (2017)

Submitted to the Sloan School of Management
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Operations Research

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

May 2022

Author .....
Sloan School of Management
April 29, 2022

Certified by .....
Andrew W. Lo
Charles E. and Susan T. Harris Professor, Sloan School of Management
Director, MIT Laboratory for Financial Engineering
Thesis Supervisor

Accepted by .....
Patrick Jaillet
Dugald C. Jackson Professor
Department of Electrical Engineering and Computer Science
Co-Director, Operations Research Center

Financial and Analytic Innovations for Therapeutic Development

by

Qingyang Xu

Submitted to the Sloan School of Management
on April 29, 2022, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy in Operations Research

Abstract

Despite groundbreaking advances in biomedicine over the past decades, the process of developing novel drug candidates from laboratory discoveries to safe and effective therapeutics approved by the Food and Drug Administration (FDA) has become longer, more expensive, and less likely to succeed. As a result, there is a widening gap in financing the clinical development of novel drug candidates, preventing potentially effective therapies from reaching the patients who are direly in need for a cure. This thesis proposes financial and data analytic innovations to address four important and challenging aspects of drug development.

We begin with an overview of the financial challenges in novel drug development and the strategies proposed to improve the financial efficiency in Part I. In Part II, we apply the "megafund" portfolio approach of financing novel drug developments to two disease areas: glioblastoma therapeutics and mRNA vaccines for emerging infectious diseases. By calibrating the simulation parameters with inputs from domain experts, we find a sharp contrast between the risk/return profiles of the two megafunds. While the megafund for glioblastoma achieves an attractive rate of return and net present value for the investors, the megafund for mRNA vaccines is unlikely to generate financial value mainly because the limited revenue of vaccine sales is insufficient to recover the significant cost of conducting late-stage clinical trials. The intrinsic limitation of the vaccine development business model motivates more cost- and time-efficient clinical trial designs discussed in Part III.

Next, in Part III, we propose a novel clinical trial design which combines Bayesian decision analysis and epidemic modeling to accelerate the clinical testing of anti-infective therapeutic candidates during a rapidly evolving epidemic outbreak. The Bayesian optimal sample size of the clinical trial decreases when the disease is more infectious and deadly, and the corresponding optimal Type I error of FDA's decision increases. In addition, we apply Bayesian decision analysis to analyze whether the clinical evidence of a controversial phase 2 clinical trial for amyotrophic lateral sclerosis justifies FDA approval, by balancing the FDA's need to limit adverse medical effects and the patients' need for expedited access to a potentially effective therapy.

In Part IV, we investigate novel machine learning models and statistical techniques to estimate key parameters of the drug development process, including the probability that the drug candidate will receive FDA approval, the duration of clinical trials, and the correlation between clinical trial outcomes. We show that there is significant bias in the machine learning models trained on the imbalanced dataset of historical drug development outcomes. We also show that debiasing the machine learning model improves the prediction accuracy and generates financial value for the drug developer.

Finally, in Part V, we analyze two social and ethical issues of the drug development process. We illustrate the success and challenges of a disruptive pricing strategy for an osteoporosis drug, including a perverse incentive of certain health plans to favorably cover drugs with higher prices in exchange for higher rebates from the drug manufacturer. We also review the ethical controversy of using the human challenge trial (HCT), in which healthy participants are actively inoculated with the pathogen, to accelerate therapeutic development for COVID-19. We call for the wider use of quantitative modeling to assess the risk/benefit tradeoff and the proactive establishment of ethical criteria so that future HCT may be conducted with minimal delay.

Thesis Supervisor: Andrew W. Lo
Title: Charles E. and Susan T. Harris Professor, Sloan School of Management
Director, MIT Laboratory for Financial Engineering

Acknowledgments

First and foremost, I am tremendously grateful to my thesis advisor, Professor Andrew Lo, for his support and guidance throughout my PhD career which helped transform me from a student into a researcher. His unique style combining technical rigor with a cogent narrative has deeply influenced my approach to research. His tireless drive towards improving healthcare and benefiting the patients has been a constant source of inspiration for me. This thesis bears witness to my endeavor of using innovative strategies to ultimately benefit the patients, an effort I hope to continue in my future career. I also thank Professor Dimitris Bertsimas and Professor Leonid Kogan for illuminating discussions which have significantly improved this thesis.

I am also grateful to my wonderful colleagues and friends at MIT Laboratory for Financial Engineering from whom I have learned a tremendous amount through our close collaborations: Shomesh Chaudhuri, Kien Wei Siah, Chi Heem Wong, Zied Ben Chaouch, Manish Singh, and Joonhyuk Cho. Valuable research assistance from Danying Xiao, Jack Zelman, Amanda Hu, Sarah Wang, Tinah Hong, and Arturo Chavez-Gehrig is gratefully acknowledged. I would also like to thank Crystal Myler, Mavanee Nealson, Kate Lyons, and Jayna Cummings for their warm support in various aspects of my research works. I have also had great pleasure learning from my wonderful collaborators, including Michael Li, Daniela Rus, Elaheh Ahmadi, Alexander Amini, Kirk Tanner, Olga Futer, and John Frishkopf.

I am very fortunate to have "co-authored" this transformative chapter of my life with many friends at MIT and beyond. I would like to thank my friends at MIT Operations Research Center and MIT Chinese Music Ensemble for so many fond memories which made my PhD experience more enjoyable than I have ever imagined. Special thanks to my students of 15.482 in the Fall semester of 2020 who have taught me many things, among which the rewarding experience of teaching.

Finally, I would like to thank my family and my girlfriend Elaine (and our cat Sushi) for their love, encouragements, unwavering support, and lots of good humor, without which none of this thesis would be possible.

Contents

List of Figures

List of Tables

Part I

Introduction

Chapter 1

Crossing the Valley of Death of Translational Biomedical Research

1.1 Background and Previous Works

It is widely believed that biomedicine is at an inflection point. The "omics" revolution in biomedical research has produced miraculous therapies for previously incurable diseases and highly safe and effective vaccines during the COVID-19 pandemic in record speed. Meanwhile, the financial efficiency of novel drug development has continued to decline. A study in 2012 found that the number of new drugs approved by the U.S. Food and Drug Administration (FDA) per billion U.S. dollars (USD) spent on translational research has halved about every 9 years since 1950 [1]. The institutional features of the drug development process, such as low probability of success [2], [3], large capital investment [4], long investment horizon [5], and high cost of capital [6] together create a financial "valley of death" [7] hindering the clinical development of novel therapeutic candidates from laboratory discoveries to live-saving therapies [8].

In the past decade, many innovations in financial engineering, clinical trial design, data analytics, and regulatory criteria have been proposed and implemented to bridge the significant funding gap of novel drug development. We review four aspects of innovations which are the most relevant to the works in this thesis and refer the readers to [8] for a comprehensive and updated review.

Financial innvoations. The financial risks of investing in novel drug development programs can be mitigated via techniques of financial engineering, such as diversification and securitization. Fernandez *et al.* [9] proposed the biomedical "megafund", a specialized financing vehicle which invests in a large portfolio of drug development programs at various clinical stages and issues both equity and securitized debt to attract private sector investors with different risk preferences. This "multiple shots on goal" approach yields a risk/return profile attractive to large institutional investors and increases the probability of developing effective therapies for presently incurable diseases. Originally proposed to finance oncology drug development, this paradigm was subsequently applied to other disease areas **Das18Pediatric Chaudhuri19Ovarian**, [10], [11] and adopted by public and private institutions such as BridgeBio Pharma and National Brain Tumor Society [12].

Clinical trial design. The costs and duration of clinical trial development can also be reduced by applying the Bayesian trial design which strikes the optimal balance between the FDA's imperative to limit the risks of approving drugs with no therapeutic efficacy or adverse effects (Type I error) and the patients' need for expedited access to effective therapies (Type II error). Traditionally, the FDA requires clinical evidence of therapeutic efficacy with p-value below 5% (one-sided hypothesis test) or 2.5% (two-sided test) in order to approve the New Drug Application. This results in prolonged clinical trials to accumulate statistical evidence. While the FDA's imperative to limit the false approval rate is justified for diseases with effective treatments, for lethal diseases such as pancreatic cancer, previous studies have shown that patients are often willing to bear a higher Type I error than 5% in exchange for lower Type II error and expedited approval of live-saving therapies [13]. Isakov *et al.* [14] proposed a Bayesian decision analysis (BDA) framework to determine the optimal significance level and clinical trial size for 30 leading causes of premature mortality in the U.S. based on the prevalence and severity of each disease. Subsequent works [13], [15]–[17] applied the BDA framework to other diseases and utilized patient survey to incorporate patient preference in the calibration of the BDA loss matrix.

Data analytics. The accumulation of clinical trial data and advances in data

analytics enabled more accurate prediction of the probability of success (PoS) that a drug candidate will receive FDA approval. Early pioneering studies [18]–[20] revealed important insights but were limited by relatively small sizes of the datasets, with fewer than 100 drugs or 500 clinical trials. Lo *et al.* [21] are the first to train machine learning models on the Citeline Informa dataset [22] (with more than 93,000 drugs and 380,000 clinical trials as of April 6, 2022) to forecast the PoS of drug approvals using drug and clinical trial features. Using this dataset, Wong *et al.* [2] proposed a novel PoS estimator of drug approvals which is robust against missing data and provided PoS estimates of vaccine clinical trials in the initial months of COVID-19 pandemic [23]. These studies led to the creation of Project ALPHA [3], a widely used benchmark of drug development success rates for 9 therapeutic areas updated each quarter. Recently, Siah *et al.* [24] organized a data science competition for drug approval prediction which attracted over 50 participating teams and generated many novel model architectures and feature engineering techniques.

Pricing. Innovative drug pricing and health insurance strategies are needed to make transformative therapies affordable to all patients. Montazerhodjat *et al.* [25] proposed the healthcare loan to finance large medical expenses. The pool of healthcare loans is financed with both equity and securitized debt and has an attractive Sharpe ratio of 4.0. While healthcare loan is an innovative insurance model, there are few works in the literature which analyze the pricing strategy of individual pharmaceutical companies for their marketed drugs and medical devices.

1.2 Thesis Contributions

The research works in this thesis present novel solutions to tackle the challenges discussed above. We summarize the contributions of each chapter which addresses a theoretical or practical challenge in four major aspects of drug development. At the time of thesis submission, previous versions of chapters 2, 4, 6, and 9 have been published as research articles in academic journals [26]–[29].

1.2.1 Financial Innovations for Funding Drug Development

The biomedical megafund [9] provides a general conceptual framework to finance novel drug development programs by attracting a wide group of investors in private sector with different levels of risk tolerance. When implementing the megafund for specific disease areas, additional domain knowledge is critical to calibrating the simulation parameters for financial analysis and ensuring that the portfolio is well diversified [10], [11], [30], [31]. The chapters in Part II apply the megafund approach to two disease areas: glioblastoma therapeutics (Chapter 2) and mRNA vaccines for emerging infectious diseases (Chapter 3).

In Chapter 2, we analyze the financial performance and social impacts of a hypothetical megafund which invests in 20 drug candidates currently under clinical development for glioblastoma (GBM) [26]. Fifteen drug candidates are eligible for GBM AGILE [32], a global adaptive trial platform which accelerates phase 2/3 testing for GBM therapies. We find that the portfolio, if properly diversified across different clinical phases and therapeutic mechanisms, generates an expected annualized rate of return of 14.9% per annum (p.a.) with standard deviation (SD) 24.3%. This risk/return profile is attractive to a wide group of investors in the private sector. Our simulations also show that at least one drug candidate will receive FDA approval in the next decade with probability 79.0%. Furthermore, biomedical expertise in selecting a well diversified portfolio is critical to generating financial value for the investors. Finally, we illustrate the synergy between the biomedical megafund and the adaptive trial platform GBM AGLE in simultaneously reducing the scientific and financial risks of developing transformative therapies for currently incurable diseases. Our analysis results have directly supported the National Brain Tumor Society to undertake its Brain Tumor Investment Fund in 2021 [12].

In Chapter 3, we investigate a hypothetical megafund which invests in 120 mRNA vaccine candidates against 11 emerging infectious diseases (EID) [33]. Unlike the GBM megafund in Chapter 2, the vaccine megafund has a negative expected rate of return -5.9% p.a. (SD 6.7%). The expected net present value (NPV) is $-\$9.5$

billion (SD $4.1 billion). Sensitivity analysis shows that the expected NPV remains negative unless the price per vaccine dose is above $78.00. We illustrate an intrinsic limitation of the business model of parallel vaccine development, namely that the revenue generated by approved vaccines is insufficient to cover the significant costs of clinical trials, which account for 94% of the total investments. However, the approved vaccines in the megafund portfolio can prevent 31 EID outbreaks on average (SD 13) in the next two decades, and the tremendous social benefits are not captured by the financial analysis. Our results underscore the urgency for continued collaboration between government agencies and the private sector in creating a sustainable business model and global vaccine ecosystem to prevent future pandemics.

1.2.2 Statistical Innovations for Clinical Trial Design

In response to the COVID-19 pandemic, many countries adopted unprecedented emergency measures to expedite clinical testing and regulatory review of vaccine and anti-infective therapeutic candidates under the rationale "extraordinary times call for extraordinary measures". While expedited clinical testing may save many lives and end the pandemic sooner [34], it also undermines the scientific standard and rigor of regulatory review, causing concern in the biomedical community [35]. This dilemma can be effectively addressed by designing a rational, transparent, and flexible framework to capture the rapidly evolving circumstances and complex tradeoffs of FDA's regulatory decision. The chapters in Part III apply a Bayesian decision analysis (BDA) framework to inform regulatory decisions for anti-infective therapeutic candidates during an epidemic outbreak (Chapter 4) and the controversial new drug application (NDA) of AMX0035, a therapeutic candidate for the lethal motor neuron disease amyotrophic lateral sclerosis (ALS) (Chapter 5).

In Chapter 4, we propose a Bayesian adaptive clinical trial framework to accelerate the clinical trial development of vaccine and anti-infective therapeutic candidates during a rapid epidemic outbreak based on the infectivity and severity of the disease [27], since the standard multiyear clinical trial and regulatory review process is not conducive to saving lives and preventing infections within the course of the outbreak

[36]. Our work is the first to combine epidemiology models with the BDA framework to provide a rational, transparent, and adaptable framework for the regulators and recommend smaller clinical trial size and higher tolerable significant level when the disease is more infectious and deadly. For COVID-19 (assuming a static basic reproduction number $R_0 = 2$ and initial infection percentage of 0.1%), the optimal significance level is 7.1% for a clinical trial of a nonvaccine anti-infective therapeutic and 13.6% for that of a vaccine. For a dynamic R_0 decreasing from 3 to 1.5, the corresponding values are 14.4% and 26.4%, respectively. In general, the Bayesian optimal significance levels are higher than the standard value of 5% required by the FDA, which reflects the urgent social imperative to avoid any false rejection or delayed approval of a potentially effective anti-infective therapy.

In Chapter 5, we apply the BDA framework to analyze the NDA of AMX0035,a novel therapeutic candidate for ALS [37]. The NDA is controversial since AMX0035 has not completed its phase 3 clinical trial while its phase 2 trial showed therapeutic effects (with p-value 0.03) in slowing the ALS disease progression on a relatively small number of 137 patients [38], [39]. Our BDA framework strikes the optimal balance in the tradeoff between the FDA's need to limit potential adverse effects (Type I error) and the ALS patients' need for expedited access to a potentially effective therapy (Type II error), evidenced by over 50,000 signatures from the ALS community calling for FDA approval. By calibrating the disease burdens of medical adverse effects and ALS, we find that BDA-optimal Type I error for approving AMX0035 is higher than the p-value of 3% reported in the phase 2 trial provided that the probability of the therapy being effective is at least 30%. Our recommendation for FDA approval is robust against a wide range of assumed values of BDA model parameters.

1.2.3 Data Analytics for Drug Development Forecasting

Three key parameters which influence the financial value of the biomedical megafund are the probability of success (PoS) of drug approval, the duration of clinical trials, and the Pearson correlations between clinical trial outcomes due to similar diseases or therapeutic mechanisms. The chapters in Part IV apply novel machine learning

and statistical inference methods to estimate these parameters.

In Chapter 6, we identify and mitigate the algorithmic bias of machine learning models to predict drug approval outcomes from phase 2 clinical trial results [28]. Machine learning models have increasingly been applied to predict drug development outcomes based on the intermediary clinical trial results [21], [24], [40]. However, the prediction accuracy is limited by the significant bias in the historical data in the form of imbalanced distributions of drug approval outcomes and drug features. For outcome labels, only 11.8% of all drugs in our dataset are approved by the FDA. In the input feature space, imbalance occurs where many drugs have similar properties (e.g., "me too" or repurposed drugs). The prediction model trained on the imbalanced dataset is likely to have a large algorithmic bias, measured by the variance of prediction accuracy across different subgroups in the dataset [41].

To address the bias due to data imbalance, we instantiate the Debiasing Variational Autoencoder (DB-VAE) [42], a state-of-the-art model for automated debiasing which simultaneously identifies the imbalance in input features and output labels and mitigates the bias in the model's predictions. We find that the debiased model improves the prediction performance with higher true positive rates and F_1 scores than their un-debiased counterparts. We also show that debiasing improves the net present value of late-stage drug development programs in six major therapeutic areas, ranging from \$763 to \$1365 million.

In Chapter 7, we study the key factors which impact the duration of a clinical trial, using both traditional statistical methods and novel machine learning models of survival analysis [43]. We find that the top three factors which influence the trial duration are the therapeutic area, the type of clinical trial sponsor, and the clinical trial phase. In particular, clinical trials for oncology indications last the longest on average while trials for metabolic indications the shortest. We also find that trials sponsored by pharmaceutical companies are shorter than those sponsored by academic and governmental medical centers. Phase 1 trials are the shortest on average, while the hybrid phase 1/2 trials are the longest. Our results call for the wider use of novel trial designs and greater public-private partnership in order to expedite the

trial duration.

In Chapter 8, we estimate the Pearson's correlations between clinical trial outcomes. Correlations are the key parameters which determines the volatility of the biomedical megafund portfolio [9], [10], [44] and are induced due to biomedical factors such as common disease or therapeutic mechanisms. Previous studies [11], [26], [30] use heuristic assessments of biomedical experts to estimate correlations, while no data-driven estimation methods have been proposed. We address this open problem by applying rigorous statistical inference techniques of intra-class coefficient (ICC) [45], [46] and generalized estimating equations (GEE) [47]–[49] on the Citeline Informa dataset. While the non-parametric ICC estimator does not yield statistically significant correlations, the parametric GEE estimator yields positive and statistically significant correlations in all therapeutic areas ranging from 2.0% (central nervous system) to 7.3% (metabolic).

1.2.4 Social and Ethical Aspects of Drug Development

In addition to the financial risks, the drug development process often poses risks on the patients' health either directly in a clinical trial due to adverse medical effects or indirectly post FDA approval due to high drug prices which prevent the disadvantaged patients from accessing live-saving therapies. The ethical and social issues associated with the drug development process must be addressed to ensure that the risks and benefits of drug development are shared across all patients in an equitable manner. The chapters in Part V delve into the ethical and practical implications of a disruptive drug pricing strategy (Chapter 9) and of conducting human challenge trials to expedite vaccine development for infectious diseases (Chapter 10).

In Chapter 9, we examine the success and challenges of the disruptive pricing strategy of *abaloparatide*, an osteoporosis drug launched in 2017 with 45% lower list price than its main competitor [29]. This disruptive strategy allowed *abaloparatide* to rapidly gain access to this market, achieve a quarterly growth of 8.5% in patient volume, and surpass its revenue guidance in 2018. However, it also faces two institutional challenges from the Medicare Part D (MPD) insurance system, which covers

50% of its patient population. First, the low MPD coverage rate of 67% is the result of a perverse incentive for certain health plans to selectively reimburse drugs with higher list prices in exchange for higher rebates from the drug manufacturer. In addition, the coverage gap in MPD leads to high out-of-pocket costs for the patients despite the lower list price, causing 50% of the patients to discontinue the treatment before the prescribed period of 18 months. Overall, we find that this pricing strategy is sustainable for the drug manufacturer, beneficial for the patient, and may have potential applications in other therapeutic areas.

Finally, in Chapter 10, we review the ethical controversies surrounding the use of controlled human challenge trials (HCT), in which participants are actively inoculated with the pathogen, to accelerate vaccine development for infectious diseases such as COVID-19. We argue that in addition to principle-driven ethical arguments, regulators should also use data-driven modeling of vaccine development during the pandemic (such as [34]) in order to rigorously quantify the risks and benefits of an HCT under different hypothetical scenarios. An HCT may be ethically conducted only if its risk-benefit tradeoff is favorable and sufficiently robust under perturbations of the model parameters. In addition, regulatory agencies and stakeholders should proactively establish the ethical criteria so that future HCTs can be initiated with minimal delay if deemed ethical.

Part II

Financial Innovations for Funding Drug Development

Chapter 2

Financing therapeutic development for glioblastoma

Development of curative treatments for glioblastoma (GBM) has been stagnant in recent decades largely because of significant financial risks. A portfolio-based strategy for the parallel discovery of breakthrough therapies can effectively reduce the financial risks of potentially transformative clinical trials for GBM. Using estimates from domain experts at the National Brain Tumor Society (NBTS), we analyze the performance of a portfolio of 20 assets being developed for GBM, diversified across different development phases and therapeutic mechanisms. We find that the portfolio generates a 14.9% expected annualized rate of return. By incorporating the adaptive trial platform GBM AGILE in our simulations, we show that at least one drug candidate in the portfolio will receive U.S. Food and Drug Administration (FDA) approval with a probability of 79.0% in the next decade.[1]

[1]Joint work with Kien Wei Siah, Kirk Tanner, Olga Futer, John J. Frishkopf, and Andrew W. Lo. An early version of chapter was published in *Drug Discovery Today* [26]. Valuable feedbacks from David Aron, Meredith Buxton, Tim Cloughesy, and Rachel Rosenstein-Sisson are gratefully acknowledged.

2.1 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 2017 study reported that GBM accounted for 47.1% of primary malignant brain tumor incidence in the U.S., while its five-year relative survival rate was only 5.5%, significantly worse than the survival rate for all malignant brain and central nervous system tumors combined, 34.9% [50]. Under the current standard of care, consisting of maximal surgical resection followed by chemoradiation [51], about 70% of GBM patients experience recurrence within one year of diagnosis, and the median survival time is merely 14.4 months [52].

Developing curative treatments for GBM is an urgent social imperative. Nevertheless, it is financially risky, due to a long investment horizon and a low probability of success. The financial risks of GBM drug development could be mitigated via the "multiple shots on goal" strategy of a "megafund" vehicle [9]. 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 made attractive to many private sector investors. Furthermore, the parallel discovery approach greatly increases the chance of producing breakthrough therapies for presently incurable diseases.

The megafund vehicle was originally proposed to finance translational research in oncology [9], and it was subsequently adapted to specific disease areas such as orphan diseases [10], Alzheimer's disease [11], and ovarian cancer [30]. It is currently under consideration as a financing vehicle by the National Brain Tumor Society (NBTS), the largest nonprofit organization in the U.S. dedicated to advancing innovative treatments of brain tumors.

In this study, we demonstrate the viability of applying the megafund vehicle to finance drug development programs for GBM. 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 diversify-

ing the portfolio across different stages of development and therapeutic mechanisms makes the risk/return profile attractive to a large group of investors in the private sector. Furthermore, we demonstrate the synergy between the megafund and the platform clinical trial program Glioblastoma Adaptive Global Innovative Learning Environment (GBM AGILE) [53], [54] in simultaneously reducing the scientific and financial risks of developing innovative GBM therapies.

2.2 Methods

In this study, we quantitatively demonstrate the synergy between a GBM megafund portfolio and an adaptive clinical trial platform in expediting the drug development process of GBM while achieving a risk/return profile attractive to a wide group of financial investors. To this end, we analyze a hypothetical portfolio of 20 real-world GBM clinical trials (Table 2.1), selected by the NBTS network of experts in GBM drug development. By combining their domain expertise with an extensive literature review, we estimate the probability of success of each drug candidate, the correlations between clinical trial outcomes, and the revenue of a transformative GBM therapy. We also include the adaptive clinical trial platform GBM AGILE in our simulations of clinical trial developments of the portfolio's assets. The detailed description of our assumptions and methodology is provided in Appendix A.1.

Table 2.1: Hypothetical GBM megafund portfolio of brain cancer therapeutics.

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

We assume that projects targeting pediatric patients are eligible for priority review vouchers. Abbreviations: ATM, ataxia-telangiectasia mutated; ATR, ataxia telangiectasia and Rad3-related protein; BBB, blood–brain barrier; DDR, DNA damage repair; DE, devices; DNA-PK, DNAdependent protein kinase; DRD2, dopamine receptor D2; EGFR, epidermal growth factor receptor; FGFR, fibroblast growth factor receptor; GA, eligibility for GBM AGILE; HGG, high-grade gliomas; IMM, immunotherapy; LPCAT1, lysophosphatidylcholine acyltransferase 1; ODS, eligibility for orphan drug status; PM, precision medicine; PP, target pediatric patients; TM, tumor metabolism; TT, transformative treatment; uMGMT, unmethylated O6-methylguanine DNA methyltransferase.

2.3 Results

The performance statistics of GBM megafund simulations is summarized in Table 2.2. The mixed-stage portfolio (row 1 in Table 2.2) illustrates the performance of the fund under the baseline assumptions. We find that its expected annualized return of 14.9% outperforms similar megafund portfolios for Alzheimer's disease [11] and ovarian cancer [30] and thus, it may attract a wide group of private sector investors. Its net present value (NPV) is $82 million, indicating that the megafund is likely to generate financial value for investors.

On the other hand, this portfolio has a 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, more than two therapies financed by the megafund will receive FDA approval. There is a 79.0% 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.

To analyze the robustness of the simulation results against each model assumption, we perform sensitivity analyses on the acquisition strategy, the correlation structure, the added value of biomedical expertise, as well as the effect of inclusion of portfolio assets in the GBM AGILE platform trial.

2.3.1 Early-Stage vs. Mixed-Stage

The performance of the portfolio hinges on its diversification. To gauge the effect of diversifying the assets across different stages of development, we simulate a comparison portfolio (row 2 in Table 2.2) with the same drug development programs, but acquiring all its assets at their preclinical stage. Preclinical acquisition requires an average investment of only $673 million, much lower than the $1.037 billion of the mixed-stage portfolio, since market valuations are based on lower probabilities of success and longer investment horizons. However, a lack of diversification across different development stages significantly increases the risk that no therapy in the

portfolio will receive FDA approval, leading to a 3.4 percentage point decrease in its expected annualized return, an 11.4 and 12.7 percentage point increase in its probabilities of loss and wipeout, respectively, and a negative NPV. It also delays the expected time until the first approved drug by 3.2 years. We conclude that, to ensure an attractive risk/return profile of the megafund, it is critical to structure the portfolio with assets acquired in different stages of development.

2.3.2 Qualitative Correlation vs. Equicorrelation

The volatility of the portfolio is largely determined by the correlation structure of the portfolio's drug development programs. It is reasonable to expect that drugs with similar therapeutic mechanisms are highly correlated, leading to greater volatility. We simulate portfolios where the correlation ρ between any two distinct assets is the same, and set to 0, 10%, 40% and 80%, respectively (rows 3 to 6 in Table 2.2). We find that the expected annual return decreases for higher correlation, while all risk measures (probability of loss and wipeout, volatility of annual return) increase.

The correlation structure of our mixed-stage portfolio is based on the qualitative assessment of program similarity by domain experts (see Appendix A.3). Although certain groups of drugs in the portfolio are highly correlated due to similar therapeutic mechanisms, diversification across different therapeutic mechanisms can lower the overall correlation to the equivalent of a uniform correlation between 10% and 40%.

2.3.3 Skill and Access Factor

There is an intrinsic limitation on GBM megafund performance due to scientific challenges of developing curative treatments for GBM. The financial viability of the GBM megafund relies on the assumption that biomedical experts are skilled at identifying promising drug candidates. This boost in probability of success is modeled by the skill and access factor α_{skill} (which is set to 1.25). Reducing α_{skill} to 1—implying no incremental improvement in the probability of success above the industry average—decreases expected annualized return by 2.0 percentage points and increases

the probabilities of loss and wipeout by 3.5 and 3.8 percentage points, respectively (row 7 in Table 2.2). The expected NPV also decreases to less than one-fourth of its original value. The sensitivity of megafund performance to α_{skill} reveals the critical importance of biomedical expertise in active management of the portfolio.

2.3.4 Transformative Factor

Our simulation also assumes that domain experts can identify potentially transformative therapies that, once approved, will become the standard of care for GBM, thus generating higher revenue than the palliative therapies. This boost in future revenue for transformative therapies is modeled by the transformative factor α_{trans}, which is set to 2. Reducing α_{trans} to 1 yields a 6.2 percentage point decrease in the expected annualized return, and a 5.9 percentage point increase in the probability of loss (row 8 in Table 2.2). Furthermore, the expected NPV becomes negative, which indicates that the ability to identify transformative therapies significantly impacts the market valuation of the portfolio.

2.3.5 Market Penetration Rate

A key factor in determining the revenue of the GBM megafund is the market penetration rate of an FDA-approved therapy, i.e. the proportion of target patient population who will receive this therapy once it enters the market. Our baseline model assumes that the maximum market penetration rate of any approved asset, r_{mkt}, is 20%. This estimate is likely conservative, since currently no curative treatment of GBM is available. Once a transformative therapy receives FDA approval, however, it is expected to become the new standard of care and may acquire a market share well above 20%. Boosting r_{mkt} to 30% increases the expected annualized return by 3.2 percentage points and doubles the expected NPV (row 9 in Table 2.2). However, reducing r_{mkt} to 10% decreases the expected annualized return by more than half, and the expected NPV becomes negative (row 10 in Table 2.2). The impact of the market penetration rate on the expected return illustrates the significant potential for the biopharma in-

dustry to develop high-risk yet truly transformative therapies for presently incurable diseases such as GBM.

2.3.6 Quantiles of Annualized Return and NPV

We report the 25%, 50% and 75% quantiles of the annualized return and NPV in Table 2.3 to measure the volatility of the megafund portfolio. We note that, while the median of annualized return (column 4) closely tracks its mean value (column 1), the median NPV (column 9) is significantly lower than its mean value (column 6), and is negative for all simulated portfolios except for those with zero correlation (row 3), the portfolio with minimum volatility. The histogram of the NPV of the baseline portfolio (Figure 2-1) reveals a bimodal distribution with a heavy right tail. The probability of a negative NPV is 54.9%, while the probabilities of an NPV above $100 million and $1 billion are 40.4% and 12.7%, respectively. The GBM megafund portfolio necessarily involves large volatility, reflecting both the inherent scientific challenge to develop an effective therapy for GBM, but also the considerable revenue once it is approved.

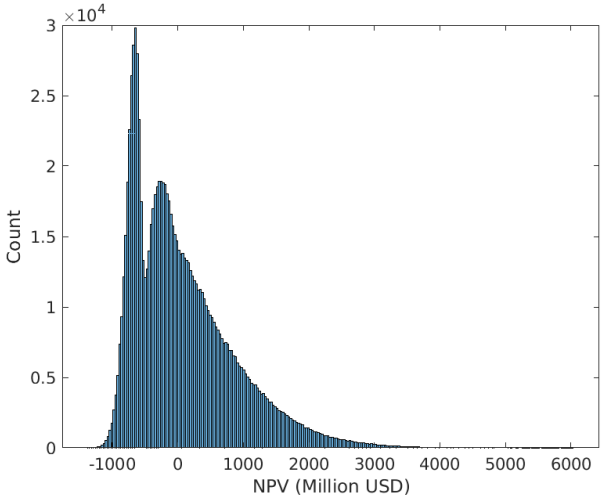


Figure 2-1: Histogram of net present value (NPV) of the baseline portfolio.

Table 2.2: Performance of GBM megafund portfolio.

Portfolio	$\mathbb{E}[R_a]$ (%)	SD$[R_a]$ (%)	$\mathbb{E}$[NPV] (M$)	SD[NPV] (M$)	$\mathbb{E}[N_a]$	SD$[N_a]$	$\mathbb{E}[T_a]$ (years)	SD$[T_a]$ (years)	PoL (%)	PoW (%)
Baseline	14.9	24.3	82	776	2.2	2.0	8.3	1.7	25.7	21.0
Preclinical	11.5	26.3	−20	399	1.5	1.6	11.5	0.9	37.1	33.7
$\rho = 0\%$	17.4	18.6	82	576	2.2	1.4	8.2	1.6	14.3	9.5
$\rho = 10\%$	16.1	21.4	82	670	2.2	1.7	8.2	1.6	19.8	14.8
$\rho = 40\%$	12.1	29.1	82	955	2.2	2.5	8.2	1.6	35.1	30.4
$\rho = 80\%$	4.4	42.7	84	1416	2.2	3.8	8.1	1.5	56.6	53.6
$\alpha_{skill} = 1$	12.9	25.0	19	741	2.0	1.9	8.3	1.7	29.2	24.8
$\alpha_{trans} = 1$	8.7	19.3	−61	434	2.2	2.0	8.3	1.7	31.6	21.0
$r_{mkt} = 10\%$	6.4	17.5	−94	375	2.2	2.0	8.3	1.7	32.8	21.0
$r_{mkt} = 30\%$	18.1	27.0	168	1184	2.2	2.0	8.3	1.7	24.4	21.0

$\mathbb{E}[R_a]$ denotes expected annualized return of the megafund portfolio and SD$[R_a]$ its standard deviation; NPV denotes net present value, in millions of USD; N_a and T_a denote the number of approved assets and time until the first FDA approval, respectively; PoL indicates the probability of loss (negative return) and PoW the probability of wipeout (all projects fail). In equi-correlated portfolios, ρ denotes the correlation between each pair of therapies; p.a. indicates per annum. The baseline portfolio assumes skill and access factor $\alpha_{skill} = 1.25$, transformative factor $\alpha_{trans} = 2$, market penetration $r_{mkt} = 20\%$ and correlation derived from estimates by NBTS network of experts. Results are computed with 1 million Monte Carlo simulations.

Table 2.3: Quantiles (25%, 50%, 75%) of annualized return (R_a) and net present value (NPV).

Portfolio	$\mathbb{E}[R_a]$ (%)	SD$[R_a]$ (%)	25% Qt. (%)	50% Qt. (%)	75% Qt. (%)	$\mathbb{E}$[NPV] (M$)	SD[NPV] (M$)	25% Qt. (M$)	50% Qt. (M$)	75% Qt. (M$)
Baseline	14.9	24.3	−0.7	14.4	30.7	82	776	−551	−98	499
Preclinical	11.5	26.3	−14.2	10.5	28.9	−20	399	−311	−128	179
$\rho = 0\%$	17.4	18.6	5.4	17.1	29.4	82	576	−345	17	441
$\rho = 10\%$	16.1	21.4	3.0	15.9	30.0	82	670	−424	−34	470
$\rho = 40\%$	12.1	29.1	−14.2	10.2	30.6	82	955	−622	−214	492
$\rho = 80\%$	4.4	42.7	−26.6	−14.2	26.4	84	1416	−677	−547	289
$\alpha_{skill} = 1$	12.9	25.0	−4.9	12.4	29.1	19	741	−586	−158	407
$\alpha_{trans} = 1$	8.7	19.3	−3.1	9.0	21.5	−61	434	−392	−147	180
$r_{mkt} = 10\%$	6.4	17.5	−4.0	6.8	17.7	−94	375	−369	−170	106
$r_{mkt} = 30\%$	18.1	27.0	0.8	16.9	35.7	168	1184	−806	−115	809

$\mathbb{E}[R_a]$ denotes expected annualized return of the megafund portfolio and SD$[R_a]$ its standard deviation; NPV denotes net present value, in millions of US$; Qt denotes quantile. The quantiles show large deviations in both annualized return and NPV from their mean values. In particular, the median (50% Qt.) NPV is negative for all megafund portfolios except the one with zero correlation (row 3). We also note that the significant risks of financial loss are compensated by the attractive annualized return and NPV values at the 75% Qt. Results are computed with 1 million Monte Carlo simulations.

2.4 Impact of GBM AGILE

The GBM megafund and GBM AGILE share the same "multiple shots on goal" strategy and have complementary goals: the former facilitates the financing of drug development programs, while the latter expedites the clinical trial process. The simulated megafund portfolio includes 15 out of its 20 assets eligible for GBM AGILE. Through detailed modeling of the GBM AGILE platform, we find that the combination of these two novel models generates significant synergy, accelerating the development of innovative therapeutics for GBM. The impact of GBM AGILE on the megafund performance is summarized in Table 2.4.

2.4.1 Probability of Inclusion

Each eligible asset in the megafund portfolio, upon successful completion of its phase 1 trial, has a probability p_{inc} to be included in stage 1 of GBM AGILE. The decision to include an asset is based on multiple factors, including its phase 1 results, the current number of experimental arms in the platform, and the expertise of the NBTS network of experts in selecting drug candidates with promising enrichment biomarkers. Our baseline model assumes a $p_{inc} = 33\%$. Varying p_{inc} from 0 to 66% (rows 2 to 5 in Table 2.4), we find that the expected annualized return increases by 7.4 percentage points, the probabilities of loss and wipeout decreases by 5.6 and 3.8 percentage points, respectively, and the expected time until first approval shortens by one year. In the absence of GBM AGILE ($p_{inc} = 0$), the expected NPV of the portfolio becomes negative, indicating that the megafund will not generate financial value for the investors. Having more assets included in the GBM AGILE platform boosts the portfolio's annualized return and NPV, reduces its risks, and accelerates the advent of transformative GBM therapies.

2.4.2 Monthly Patient Accrual

Another crucial factor of GBM AGILE is the monthly patient accrual rate into the platform. A lower accrual rate delays the completion of stage 1 and 2 investigations

and lowers the NPV due to longer investment horizons. We assume an accrual rate $v_{mon} = 30$ patients per month in our baseline model. This is a relatively conservative estimate, since GBM AGILE may potentially recruit patients in the U.S., Canada, China, Europe, and Australia. Increasing v_{mon} to 40 and 50 patients per month (rows 7 to 8 in Table 2.4) increases the expected NPV of the megafund from \$82 million to \$114 million and \$134 million, respectively, and shortens the expected time until first approval from 8.3 years to 8.0 and 7.8 years, respectively. On the other hand, reducing v_{mon} to 20 patients per month (rows 6 in Table 2.4) lowers the expected NPV to \$31 million, and delays the expected time until first approval from 8.3 to 8.9 years. The success of the GBM megafund hinges critically on the steady accrual of new patients to support the speedy completion of stages 1 and 2 of GBM AGILE.

Table 2.4: Impact of GBM AGILE on megafund portfolio performance.

Portfolio[2]	$\mathbb{E}[R_a]$ (%)	SD$[R_a]$ (%)	$\mathbb{E}$[NPV] (M$)	SD[NPV] (M$)	$\mathbb{E}[N_a]$	SD$[N_a]$	$\mathbb{E}[T_a]$ (years)	SD$[T_a]$ (years)	PoL (%)	PoW (%)
Baseline	14.9	24.3	82	776	2.2	2.0	8.3	1.7	25.7	21.0
$p_{inc} = 0\%$	11.5	22.3	−40	712	2.2	2.0	8.9	1.3	28.6	23.0
$p_{inc} = 16\%$	13.1	23.1	20	743	2.2	2.0	8.6	1.5	27.1	22.0
$p_{inc} = 50\%$	16.9	25.6	147	809	2.2	2.0	8.1	1.7	24.2	20.0
$p_{inc} = 66\%$	18.9	27.0	205	835	2.3	2.0	7.9	1.7	23.0	19.2
$v_{mon} = 20$	14.5	23.6	31	713	2.2	2.0	8.9	1.4	25.5	20.9
$v_{mon} = 40$	15.1	24.6	114	817	2.2	2.0	8.0	1.9	25.6	20.9
$v_{mon} = 50$	15.3	24.8	134	843	2.2	2.0	7.8	2.1	25.5	20.9

$\mathbb{E}[R_a]$ denotes expected annualized return of the megafund portfolio and SD$[R_a]$ its standard deviation; NPV denotes net present value, in millions of US$; Qt denotes quantile. The quantiles show large deviations in both annualized return and NPV from their mean values. In particular, the median (50% Qt.) NPV is negative for all megafund portfolios except the one with zero correlation (row 3). We also note that the significant risks of financial loss are compensated by the attractive annualized return and NPV values at the 75% Qt. Results are computed with 1 million Monte Carlo simulations.

2.5 Discussion

The development of transformative therapeutics for GBM has been largely unsuccessful due not only to the inherent scientific challenges of development, but also due to the significant financial risks of investing in early-stage clinical programs. The performance of a GBM megafund may attract a wide group of investors from both the public and private sectors, especially if it has a suitably diversified portfolio managed by domain experts.

In addition, the use of the novel GBM AGILE platform generates significant synergy with the megafund. Inclusion of portfolio assets in the platform boosts its annualized return and NPV, reduces its risks, and expedites the ultimate delivery of transformative GBM therapies, making it more attractive to private sector 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 exists 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 $1.5 billion 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 [25]. Under a tight budget constraint, it may also be necessary to acquire drug development programs dynamically, liquidating some projects during intermediary development to fund more promising ones. Our simulation results can be regarded as an upper bound on the performance of a GBM megafund in practice.

Finally, the financial performance of the megafund may be further improved via portfolio optimization strategies[3]. We discuss an ongoing effort to optimize the megafund portfolio via the techniques of option pricing in Section A.7.

[3]We thank Dimitris Bertsimas and Leonid Kogan for suggesting viable optimization strategies.

2.6 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 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 improve development outcomes, the risk/return profile of such a megafund should interest many private sector investors.

Chapter 3

Financing mRNA vaccine development for emerging infectious diseases

We analyze the financial performance of a hypothetical vaccine megafund based on mRNA technology. The portfolio consists of 120 mRNA vaccine candidates in the preclinical stage targeting 11 emerging infectious diseases. We calibrate the simulation parameters with input from domain experts in mRNA technology and an extensive literature review. We find that the megafund portfolio generates an average annualized return on investment of -6.0% per annum and a net present value of $-\$9.5$ billion. Clinical trial costs account for 94% of the total investment, with manufacturing costs accounting for only 6%. Sensitivity analysis reveals that the most important factor of financial performance is the price per dose, while the increased probability of success due to mRNA technology, adjusting the size of the megafund portfolio, and the possibility of conducting human challenge trials do not significantly improve financial performance.[1]

[1]Joint work with Zied Ben Chaouch, Michael Li, Dimitris Bertsimas, Jacob Becraft, Tasuku Kitada, Joseph Barberio, Kevin Shi, and Andrew W. Lo. Valuable discussions and feedback from Regina Dugan and Ken Gabriel are gratefully acknowledged.

3.1 Introduction

The incalculable human, social, and financial costs of the COVID-19 pandemic has created the collective imperative to prepare for the next pandemic by proactively engaging in the research and development (R&D) of novel vaccines against emergent infectious diseases (EIDs). A notable example of such an effort is the Coalition for Epidemic Preparedness Innovations (CEPI), which has created a portfolio of 14 vaccine candidates targeting COVID-19 and six other priority EIDs as of January 11, 2022 [55].

Vaccine R&D has also undergone a revolution during the pandemic—exemplified by the messenger RNA (mRNA) technology—which has demonstrated robust safety, high efficacy, and unprecedented speed of clinical development (see [56] for an updated review). This technology has the potential to significantly reduce the cost and duration of vaccine R&D, enabling much more rapid responses to future EIDs. It is also particularly suited to the portfolio-based approach of CEPI, since different mRNA vaccine candidates may share the same resources and facilities of preclinical animal studies, clinical testing, and post-approval manufacturing and delivery.

An important challenge to the portfolio-based model of mRNA vaccine development is the lack of sufficient and sustainable funding to support the vaccine development pipeline over the extended period (typically over multiple years) from preclinical research to FDA approval, an issue known as the "valley of death" in translational biomedical research [7]. Governments, international agencies, and non-governmental organizations have contributed significantly to create a sizeable portfolio of vaccine candidates, but their efforts have nevertheless fallen short. However, the private sector may provide the investment needed to finance the vaccine R&D pipeline, provided that the vaccine portfolio can generate financial returns for its investors.

In this paper, we simulate a hypothetical vaccine megafund with a large portfolio of 120 mRNA vaccine candidates targeting 11 EIDs and ask whether the risk/return profile of the megafund is attractive to private-sector investors. To address the complexity of simulating the vaccine development process, we calibrate the simulation

parameters with input from domain experts in mRNA technology and an extensive literature review. We illustrate the key factors affecting the financial performance of the vaccine megafund, and discuss potential solutions to improve its financial returns to investors.

3.2 Literature Review

The lack of sufficient funding for novel translational R&D is due to several institutional features of drug development, including a low probability of success (PoS), a long investment horizon, high clinical trial costs, and a high cost of capital [8]. To address this funding challenge, Fernandez *et al.* [9] proposed a novel financing vehicle, the biomedical megafund, which invests in a sizable portfolio of drug candidates diversified across different clinical stages and therapeutic areas. Using techniques from financial engineering, the authors show that the risk/return profile of the megafund is attractive to a wide group of investors. The megafund model was subsequently applied to various disease areas, including orphan diseases [10], Alzheimer's disease [11], pediatric cancer [31], ovarian cancer [30], glioblastoma [26] and vaccines against EIDs [33]. It is currently being undertaken by the National Brain Tumor Society (NBTS) to finance novel drug candidates to treat glioblastoma [12].

In the simulation analysis of [33], the financial performance of a vaccine megafund is extremely unfavorable to for-profit investors, with expected annualized return of -61% and standard deviation (SD) of 4%. Multiple factors lead to this negative financial return, including a low probability of success (PoS) of vaccine trials, high clinical trial costs, and limited revenue from vaccine sales. Based on these findings, the authors propose several strategies to finance the vaccine megafund, such as a price increase, public sector funding, and a novel subscription model in which subscribers would pay annual fees for priority access to the vaccine during future outbreaks.

In this paper, we extend the previous work [33] in several ways. First, previous work simulated stochastic vaccine trial outcomes, but used the expected annual profit for approved vaccines. We implement a more realistic simulation framework in which

the entire pipeline of vaccine development and manufacturing is simulated under stochastic occurrence of EID outbreaks. The uncertainty in future EID outbreaks increases the variance of megafund cash flows, and is critical to its risk/return profile. In addition, we use improved PoS estimates of mRNA vaccines to adjust the cash flows of the megafund, and calibrate the cost structure of mRNA vaccine manufacturing with input from domain experts and an updated literature review. Finally, while [33] focused on annualized return, we systematically investigate a wide spectrum of performance metrics, such as the net present value and the number of EID outbreaks prevented, and provide a detailed breakdown of megafund investment to identify the main bottlenecks in financial performance.

The financial performance of the megafund also hinges on the scientific and business expertise of fund managers to select promising drug candidates and diversify the portfolio [26]. For a real-world vaccine portfolio such as CEPI's, active portfolio management is especially important given the budget constraints to undertake a limited number of vaccine candidates. Gouglas and Marsh [57] apply multi-criteria decision analysis to select promising vaccine candidates for the CEPI portfolio in the context of multiple trade-offs and heterogenous stakeholder preferences. In a subsequent study [58], the authors apply portfolio decision analysis to optimize the investment of CEPI in 16 vaccine technology platforms. Another recent study [59] analyzed the optimal investment strategy of vaccine manufacturing capacity for countries with different socioeconomic characteristics.

While we recognize the importance of portfolio management in improving the financial performance of the vaccine megafund, we do not impose exogenous budget constraints or perform portfolio optimization in our simulation analysis, since our goal is to understand the relationships between the investment and revenue of the vaccine megafund and endogenous factors, such as the improved probability of success of mRNA vaccines, the cost structure of mRNA vaccine manufacturing, and the possibility of conducting human challenge trials to expedite the vaccine development process.

3.3 Methods

3.3.1 Vaccine megafund portfolio

We simulate the financial performance of a large portfolio of vaccine candidates shown in Table 3.1. The portfolio structure and probability of outbreak P_a of each EID are adapted from [33]. We also include 10 vaccine candidates which target "disease X", the next pandemic of an unknown pathogen, in accordance with the updated CEPI portfolio [55]. We assume that disease X has an annual probability of outbreak $P_a = 1\%$ and the number of infected cases is 400 million, close to that of COVID-19.

Table 3.1: Portfolio for vaccine megafund simulation.

Infectious Disease	N_{vac}	$P_a(\%)$	n_I
Disease X	10	1.0	400,000,000
Chikungunya	16	10.8	523,600
Zika Virus	18	4.3	500,062
Lassa Fever	7	100.0	300,000
Rift Valley Fever	3	10.5	79,414
SARS-CoV-1	2	7.1	8,098
West Nile Virus	23	10.0	500
MERS-CoV	8	40.0	436
Crimean-Congo Hemo. Fever	7	12.5	320
Nipah Virus	20	15.8	136
Marburg Virus	6	12.0	75

N_{vac} denotes the number of vaccine candidates targeting each emerging infectious disease (EID); P_a denotes the annual probability of outbreak; n_I denotes the average number of infected cases.

3.3.2 Vaccine clinical trials

We use the simulation framework in [26] to model correlated phase transitions of vaccine clinical trial developments. The assumed values of the simulation parameters of a vaccine clinical trial are summarized in Table 3.2. The simulated trial outcomes depend on two critical sets of parameters. First, the probability of success (PoS) for each phase transition in the clinical development process is estimated using historical industry average values [3]. In addition, since the mRNA vaccine for COVID-19 induces humoral immune protection by producing neutralizing antibodies [60], we assume that mRNA vaccines have a higher PoS for six infectious diseases whose correlate of protection is a neutralizing antibody (Chikungunya virus, SARS-CoV-1, Marburg virus, Rift Valley Fever, Nipah virus, Zika virus). To reflect the increased PoS due to mRNA technology for these diseases, we multiply their PoS by a technology factor α_{tech}. We set $\alpha_{tech} = 1.2$ in the baseline model, which reflects a 20% increase in PoS over the industry average. We do not increase the PoS for the other five diseases with cellular or unknown immune responses, including disease X. We vary α_{tech} in our sensitivity analysis to gauge the effect of increased PoS.

Table 3.2: Simulation parameters for standard clinical trials.

Parameter	PRE to P1	P1 to P2	P2 to P3	P3 to EUA	Source
PoS (%)	60.0	83.6	65.8	80.9	[2], [3]
Duration (months) Standard clinical trial	18.0	24.0	18.0	14.0	[3], [34]
Development cost (M$) Standard clinical trial	26.0	14.0	28.0	150.0	[61]
Duration (months) Human challenge trial				8.0	[34]
Development cost (M$) Human challenge trial				12.5	[34]

PRE denotes preclinical phase, P1, P2 and P3 denote phase 1, phase 2 and phase 3. We assume that vaccines receive emergency use authorization (EUA) after completing phase 3 and human challenge trial is only applicable to phase 3.

In addition, the correlations between vaccine trial outcomes have a major impact on the simulation outcomes. If two vaccine trial outcomes are highly correlated due to the same target pathogen or mechanism of action, they are more likely to simultaneously succeed or fail, which leads to greater variance in the cash flows of the mega-fund, and thus greater financial risk. Using the input of domain experts in mRNA technology, we construct a biologically motivated metric to estimate the correlations. Specifically, we use a novel distance metric $d_{i,j}$ between viruses i and j, defined as the average of similarity scores of four biological factors: taxonomy, qualitative features (e.g., type of vector, strand direction, nucleic acid topology), quantitative features (number of strands, total genome size), and the edit distance of protein sequences. The value of $d_{i,j}$ is normalized between 0 and 1, with $d_{i,j}$ closer to 0 if viruses i and j are more biologically similar, and $d_{i,j} = 0$ if they are identical. Given the values of $d_{i,j}$, a natural way to define the correlation $\rho_{i,j}$ between the outcomes of vaccine trials targeting viruses i and j is $\rho_{i,j} = 1 - d_{i,j}$.

Figure 3-1 shows the heatmap of $\rho_{i,j}$ between each pair of viruses excluding disease X (which we assume to be independent of the other viruses in order to reflect its *a priori* unknown biological properties). The correlation matrix $\rho_{i,j}$ defined this way is positive definite (PD) in our calibration, although in general it is not guaranteed to be PD, and needs to be transformed into a PD matrix by appropriate methods [62]. This metric does not specify the correlation between two vaccine trials targeting the same virus. We assume this correlation to be 0.8, which is higher than the maximum correlation across different diseases (Figure 3-1). To illustrate the impact of correlation on the financial performance, we vary the assumed values of correlation in the sensitivity analysis.

	Chikun.	SARS	MERS	Marburg	RVF	Lassa	Nipah	CCHF	WNV	Zika
Chikun.	1.00	0.30	0.30	0.37	0.27	0.39	0.38	0.29	0.38	0.33
SARS	0.30	1.00	0.58	0.32	0.21	0.25	0.28	0.26	0.29	0.28
MERS	0.30	0.58	1.00	0.33	0.20	0.25	0.28	0.26	0.29	0.28
Marburg	0.37	0.32	0.33	1.00	0.27	0.37	0.46	0.37	0.36	0.35
RVF	0.27	0.21	0.20	0.27	1.00	0.48	0.29	0.52	0.27	0.26
Lassa	0.39	0.25	0.25	0.37	0.48	1.00	0.36	0.35	0.40	0.40
Nipah	0.38	0.28	0.28	0.46	0.29	0.36	1.00	0.32	0.39	0.39
CCHF	0.29	0.26	0.26	0.37	0.52	0.35	0.32	1.00	0.29	0.28
WNV	0.38	0.29	0.29	0.36	0.27	0.40	0.39	0.29	1.00	0.64
Zika	0.33	0.28	0.28	0.35	0.26	0.40	0.39	0.28	0.64	1.00

Figure 3-1: Heatmap of correlations between vaccine candidates estimated using the distance metric $\rho_{i,j} = 1 - d_{i,j}$.

3.3.3 Vaccine manufacturing and supply chain

The cost structures of mRNA vaccine manufacturing and supply chain are key inputs to simulating the cash flows of the megafund. This information is not disclosed by the mRNA vaccine manufacturers, hence we use publicly available estimates by domain experts [63], [64] to calibrate the cost structures. The hypothetical budget of mRNA vaccine manufacturing is summarized in Table 3.3. We assume that each production line operates at a working volume of a 30L bioreactor with mRNA titer 5g/L and each vaccine dose contains 65g of mRNA (the average of the Pfizer/BioNTech and Moderna vaccines for COVID-19).

Table 3.3: Cost structure of mRNA vaccine production.

Category	Item	Unit Cost ($)	Quantity
Fixed cost	Production line (PL)	58M	1 bioreactor of 30L working volume
Variable cost	Raw materials	456.6M/(year·PL)	29,162 grams of mRNA per PL each year
	Consumables	150M/(year·PL)	
Variable cost	Labor	20/hour	113,186 labor hours per PL each year
	Quality control	10/hour	
Variable cost	Fill-and-finish	0.27/dose	10-dose vials
	Lab, utility, waste management, etc.	<1% total cost	Not modeled here

Costs of mRNA vaccine manufacturing are calibrated in [63], [64].

Using these estimates, the variable cost of producing each mRNA vaccine dose is $1.60. Assuming each EID outbreak requires 10 million vaccine doses, it takes 8.1 days to produce the mRNA needed with one production line and additional 4 to 5 weeks to perform quality control for each batch produced. The total manufacturing cost is $16 million if one uses the existing production line, and $75 million if one builds a new production line. Similarly, assuming disease X pandemic requires 1 billion vaccine doses, it takes 81.4 days to produce the mRNA needed with 10 production lines. The total cost is $1.6 billion with existing production lines, and $2.2 billion with new ones. Furthermore, we assume that the variable cost of delivering each vaccine dose in the supply chain is $1.00 (the same order of magnitude as the manufacturing cost). We make the conservative assumption on the supply chain cost due to the lack of publicly available estimates in the literature. Our simulation results show that the supply chain costs constitute 2% of total costs (Figure 3-3), so the financial performance is not sensitive to the detailed value of supply chain cost, as long as it is not higher than $1.00 by an order of magnitude.

To estimate the revenue generated by vaccine sales, we use the list prices of mRNA vaccines for COVID-19. As of October 26, 2021, thse Pfizer/BioNTech vaccine is

priced at $24.00 per dose in the U.S., and the Moderna vaccine at $15.00 per dose [65]. We assume that the price per vaccine dose is $\pi = \$20.00$. This is likely to be an underestimate, since it is below the prices of all adult vaccines (except influenza) listed in the vaccine price list of Centers for Disease Control and Prevention [66].

3.3.4 Overview of simulation framework

At the initial time $t = 0$, all vaccine candidates enter the preclinical stage. We assume that the development costs of each phase are incurred at the start of the phase. In each subsequent year from $t = 1$ to $t = T$, we simulate whether any EID outbreaks (including the disease X pandemic) will occur in year t. In the absence of any outbreaks, we develop each vaccine candidate (except the ones for "disease X") from the preclinical stage to the completion of phase 2, assuming the cost and timeline of a standard clinical trial (rows 2 and 3 of Table 3.2). We do not initiate the large-scale phase 3 clinical trial unless an outbreak has occurred, since there are no or not enough infected subjects to test the vaccine efficacy. This also reduces the clinical trial development cost compared to the estimates of [33].

If an infectious disease outbreak occurs in year t, we assume that one of the four scenarios below will occur:

1. At least one vaccine candidate targeting the disease has successfully completed a phase 3 trial during a previous outbreak of the same disease and received approval or EUA from the FDA. We manufacture the vaccine, supply to the point of distribution, and collect the revenue from the vaccine sales.

2. At least one vaccine candidate targeting the disease has successfully completed a phase 2 trial. We initiate the phase 3 clinical trial. If the phase 3 trial is successful, the vaccine receives EUA from the FDA. We manufacture and supply the vaccines, and collect the revenue from the vaccine sales.

3. At least one vaccine candidate for the epidemic is in the preclinical or phase 1 stage. We initiate an accelerated phase 1/2 trial, which costs $28 million (the

same as a standard phase 2 trial) and completes in 3 months, followed by a standard phase 3 trial, which completes in 14 months. If the phase 3 trial is successful, we manufacture and supply the vaccines, and collect the revenue.

4. No vaccine candidates for the disease have previously completed a phase 3 trial or remain in the pipeline. In this case, no cash flows are generated, since all vaccine candidates have failed in the clinical trial process.

In addition, due to the demonstrated safety and efficacy of mRNA vaccines for COVID-19, it is conceivable that human challenge trials (HCT) may be ethically justified for the mRNA vaccine candidates in the megafund portfolio, which significantly reduces the cost and duration of clinical trials. To model this possibility, we use the Bernoulli random variable HCT_i to denote whether HCT is permitted during an outbreak of disease i (with probability p_{HCT}). If $HCT_i = 1$, we use the cost and duration of HCT (rows 4 and 5 of Table 3.2) instead of the corresponding values of standard trials in scenarios 2 and 3 above. We assume $p_{HCT} = 0$ in the baseline model and illustrate the effect of p_{HCT} in the sensitivity analysis.

We simulate an investment horizon of $T = 20$ years, which includes 5 years for standard clinical trial development from the preclinical phase to the completion of a phase 2 trial, and 15 years for vaccine patent protection and market exclusivity. We compute the financial performance and social impact of the vaccine megafund at the end of the 20-year horizon.

3.4 Results

3.4.1 Baseline portfolio

The performance of the baseline portfolio is summarized in Table 3.4. We find that this portfolio has an expected annualized return $\mathbb{E}[R_a] = -6.0\%$, standard deviation $\text{SD}[Ra] = 6.7\%$ and an expected net present value (NPV) of $-\$9.5$ billion (standard error SE=\$13 million). The vaccine megafund does not generate positive financial value for the investors, since the revenue generated by the vaccine sales (\$7.5 billion

on average) is insufficient to recover the investments in clinical trial development and vaccine manufacturing ($17.7 billion on average). However, the financial value to investors does not capture the societal benefits generated by the megafund. On average, 45 infectious disease outbreaks will occur in the simulation period, 31 of which will be prevented or contained by vaccines developed from the portfolio. In addition, there is a 66% probability that vaccines in the portfolio will prevent the next "disease X pandemic" if it occurs. The lives saved and socioeconomic losses avoided by the vaccines far exceed the negative financial value of the megafund.

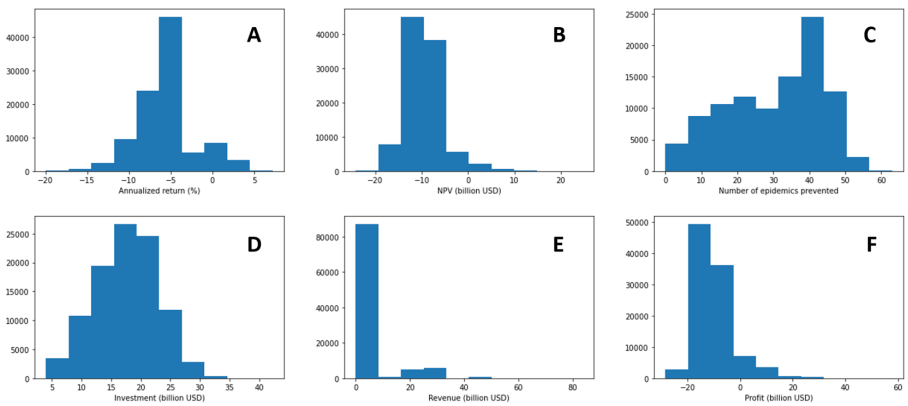
We visualize the distribution of key performance metrics of the megafund via the histograms in Figure 3-2. We find that although R_a and NPV are both negative in most simulations, there is a 9.8% probability that $R_a > 0$, and a 3.1% probability that NPV>0. In addition, the distribution of megafund investments is smooth and unimodal, while the distribution of revenue is bimodal: most of the probability mass is concentrated below $10 billion, with a small mass above $20 billion. The latter corresponds to the scenarios when disease X pandemic occurs, generating a revenue of $20 billion from vaccine sales. The bimodality of revenue leads to significant variance in the annualized return and NPV of the megafund.

To gain additional insights into the leading costs that limit the financial performance of the megafund, we present a breakdown of megafund investments in rows 9 to 15 of Table 3.4, and a corresponding pie chart in Figure 3-3. We find that the costs of clinical trial development constitute 94% of the total cost, with phase 3 trials alone accounting for 59%. The net cost of vaccine manufacturing and supply chain constitute only 6% of the total cost. Our finding reflects the "valley of death" in financing translational biomedical research [7], in which the main bottleneck is the cost of clinical trial development rather than drug manufacturing and supply. Even with more efficient vaccine manufacturing technologies and supply chain designs, the significant cost of clinical trial development still prevents the vaccine megafund from generating positive financial value to the investors.

Table 3.4: Performance of baseline portfolio computed with 100K Monte Carlo simulations.

Metric	Mean	SD	Median	25% Qt.	75% Qt.
R_a (%)	−6.0	6.7	−5.7	−7.4	−4.4
NPV (B$)	−9.5	4.1	−9.9	−12.1	−7.4
Investment (B$)	17.7	5.3	17.8	14.0	21.4
Revenue (B$)	7.5	7.7	5.8	3.4	7.0
Profit (B$)	−10.0	7.4	−11.5	−14.9	−7.5
N_{ep}	31	13	34	19	42
N_{p3}	44	23	43	26	60
N_{p2}	15	12	13	6	23
Preclinical Cost (B$)	3.1	0.0	3.1	3.1	3.1
Phase 1 Cost (B$)	1.4	0.2	1.4	1.3	1.5
Phase 2 Cost (B$)	1.7	0.6	1.7	1.2	2.2
Phase 3 Cost (B$)	10.4	4.4	10.5	7.2	13.5
Production Facility Cost (B$)	0.12	0.17	0.06	0.06	0.06
Vaccine Variable Cost (B$)	0.56	0.59	0.43	0.25	0.53
Supply Chain Cost (B$)	0.34	0.37	0.25	0.15	0.31

R_a denotes annualized return; NPV denotes net present value; N_{ep} denotes the number of epidemic outbreaks prevented by vaccines in the portfolio; N_{p3} (N_{p2}) denotes the number of vaccines which successfully complete phase 3 (phase 2) by the end of the investment horizon of 20 years. NPV is computed with an annual discount rate $r = 10\%$. NPV, investment, revenue and cost breakdown are shown in billion USD (B$). The standard deviation of preclinical trial cost is zero since the megafund invests in the preclinical trials of all 120 vaccine candidates at the initial time 0.



(A) Annualized return R_a. (B) Net present value (NPV). (C) Number of epidemics prevented N_{ep}. (D) Total investment. (E) Total revenue. (F) Net profit.

Figure 3-2: Histograms of key performance metrics of vaccine megafund.

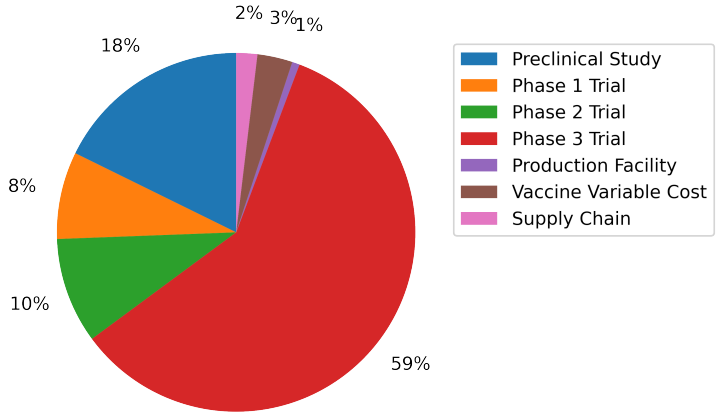


Figure 3-3: Breakdown of cost structure of the vaccine megafund.

3.4.2 Sensitivity analysis

To test the robustness of the simulation results against the assumed parameter values, we perform a sensitivity analysis for several key parameters. The results are summarized in Table 3.5 and discussed in the sections below.

Vaccine price

The price per vaccine dose π is the key driver of the financial performance. In the baseline model, we assume $\pi = \$20.00$ where both the annualized return and NPV are negative. Increasing π to \$69.00 (row 2 of Table 3.5) achieves the breakeven point for the annualized return. Increasing further to \$78.00 (row 3) achieves the breakeven point for NPV. Assuming $\pi = \$100.00$ (row 4), the megafund generates an expected annualized return of 1.9% with volatility of 7.2%, and an expected NPV of \$3.6 billion. Such a high list price of \$100.00 per vaccine dose is not unusual in the U.S. As of January 1, 2022, twelve common adult vaccines have list prices above \$100.00 in the U.S. [66]. However, these may be impossible to afford in low-to-middle income countries, and may even increase vaccine hesitancy among the affected population.

Improved probability of success of mRNA vaccines

To test whether the increased PoS of mRNA vaccines leads to improved financial performance, we multiply the PoS of vaccine trials for six diseases by the technology factor $\alpha_{tech} > 1$ to reflect the safety and efficacy of mRNA vaccines for diseases with humoral immune protection. In the baseline model, we set $\alpha_{tech} = 1.2$ (i.e., a 20% increase in PoS). Surprisingly, increasing α_{tech} from 1.0 to 1.3 (rows 5 to 7 of Table 3.5) achieves a mixed effect: the expected annualized return increased from -6.7% to -5.8%, while the expected NPV decreased from $-\$8.1$ to $-\$9.9$ billion. The reason for the mixed effect is that as we increase α_{tech} from 1.0 to 1.3, the average number of approved vaccine candidates increases from 28 to 49, while the expected investment increases from \$15.2 to \$18.4 billion. However, the expected revenue undergoes a much smaller increase from \$7.1 to \$7.6 billion, since on average only 3 additional

epidemic outbreaks are prevented by the approved vaccines (due to the stochastic occurrence of epidemic outbreaks). The smaller ratio of revenue to investment causes the annualized return to be less negative and increase, while the larger increase in investment causes the NPV to be more negative and decrease. We conclude that the higher PoS of mRNA technology alone does not generate positive financial value for the megafund unless we also reduce the clinical trial costs or raise the price of the vaccine.

Correlation

The correlation between vaccine trial outcomes measures the tendency for multiple vaccine trials to simultaneously succeed or fail due to a common target disease or mechanism of action. In the baseline model, we estimate the correlation via the novel virus distance metric $d_{i,j}$. However, we cannot simply rescale $d_{i,j}$ in the sensitivity analysis, since the resulting correlation matrix is not guaranteed to remain positive definite. Instead, we gauge the impact of correlation by assuming an equi-correlated correlation matrix in which $\rho_{i,j} = \rho$ is the same for all diseases, and vary the value of ρ from 0 (independent) to 80% (highly correlated), as shown in rows 8 to 12 in Table 3.5. As expected, we observe that higher values of ρ lead to worse financial performance, as the expected annualized return decreases from -3.5% to -11.7% and the expected NPV decreases from $-\$8.3$ to $-\$9.5$ billion. In addition, the volatility of annualized return SD$[R_a]$ dramatically increases from 2.5% to 23.6%. This shows the importance of diversity in the megafund portfolio to generate positive financial value.

Human challenge trials

If deemed ethical, an HCT can significantly reduce the cost and duration of the clinical development of vaccine candidates by testing on a smaller group of participants than traditional vaccine trials. We investigate the effect of HCTs on the megafund performance by assigning the probability p_{HCT} that HCT is allowed for each infectious disease. The baseline portfolio does not utilize HCT, i.e., $p_{HCT} = 0$. Increasing

p_{HCT} from 0 to 30% (rows 13 to 14 of Table 3.5) reduces the expected investment and increases both the annualized return and NPV, although both remain negative. We find that utilizing HCT alone is also insufficient to generate positive financial value for the megafund.

Megafund portfolio size

The parallel vaccine development strategy increases the probability that at least one vaccine candidate will be approved, but it also leads to significant costs of clinical trials. To investigate the effect of portfolio size, we multiply the number of vaccine candidates for each infectious disease by a factor γ. The baseline portfolio corresponds to $\gamma = 1$. Increasing the portfolio size by 50% ($\gamma = 1.5$, row 15 of Table 3.5) leads to worse financial performance, since the expected investment increases from \$17.7 to \$25.7 billion, while the expected revenue increases by a much smaller amount from \$7.5 to \$7.9 billion since the natural occurrence of infectious diseases remain the same. Decreasing the portfolio size by 50% ($\gamma = 0.5$, row 16 of Table 3.5) increases both expected return and NPV, though both remain negative. In addition, the average number of epidemics prevented decreases from 31 to 27, reflecting a higher loss on the society that is not captured by our financial analysis.

Table 3.5: Sensitivity analysis of key simulation parameters computed with 100K Monte Carlo simulations.

Portfolio	$\mathbb{E}[R_a]$ (%)	SD$[R_a]$ (%)	$\mathbb{E}$[NPV] (B$)	SD[NPV] (B$)	$\mathbb{E}$[[Inv] (B$)	SD[Inv] (B$)	$\mathbb{E}$[Rev] (B$)	SD[Rev] (B$)	$\mathbb{E}[N_{ep}]$	SD$[N_{ep}]$
Baseline	−6.0	6.7	−9.5	4.1	17.7	5.3	7.5	7.7	31	13
$\pi = \$69$/dose	0.0	7.1	−1.4	11.9	17.7	5.3	25.8	26.7	31	13
$\pi = \$78$/dose	0.7	7.1	0.0	13.5	17.7	5.3	29.2	30.2	31	13
$\pi = \$100$/dose	1.9	7.2	3.6	17.4	17.7	5.3	37.4	38.7	31	13
$\alpha_{tech} = 1.0$	−6.7	11.9	−8.1	4.1	15.2	5.3	7.1	7.8	28	14
$\alpha_{tech} = 1.1$	−6.2	9.1	−8.8	4.1	16.4	5.4	7.3	7.8	29	14
$\alpha_{tech} = 1.3$	−5.8	4.8	−9.9	4.1	18.4	5.1	7.6	7.7	31	13
$\rho = 0.0$	−3.5	2.5	−8.3	3.7	18.1	2.5	10.7	8.9	43	7
$\rho = 0.2$	−3.8	2.7	−8.5	4.0	18.0	3.9	10.2	8.7	41	9
$\rho = 0.4$	−4.2	4.2	−8.7	4.3	17.9	5.0	9.6	8.6	38	11
$\rho = 0.6$	−5.9	11.1	−9.0	4.6	17.8	6.0	8.7	8.3	35	14
$\rho = 0.8$	−11.7	23.6	−9.5	4.8	17.7	7.1	7.5	7.9	31	17
$p_{HCT} = 0.1$	−5.7	6.7	−8.8	4.1	16.7	5.1	7.5	7.7	31	13
$p_{HCT} = 0.3$	−5.1	6.7	−7.6	3.9	14.7	4.6	7.5	7.7	31	13
$\gamma = 0.5$	−4.1	8.9	−3.7	3.0	9.3	2.9	6.5	7.3	27	14
$\gamma = 1.5$	−7.3	5.7	−15.3	5.4	25.7	7.6	7.9	7.9	32	13

R_a denotes annualized return; NPV denotes net present value; N_{ep} denotes the number of epidemic outbreaks prevented by vaccines in the portfolio; N_{p3} (N_{p2}) denotes the number of vaccines which successfully complete phase 3 (phase 2) by the end of the investment horizon of 20 years. NPV is computed with an annual discount rate $r = 10\%$.

3.5 Discussion

Our analysis illustrates two major obstacles in financing novel vaccine development. First, the annual demand of vaccines is mainly determined by the natural occurrence of infectious disease outbreak. This limits the revenue generated by the approved vaccines, unless we increase the list price to $78.00 per dose. With such a high list price, local governments and populations may not be able to afford the vaccines, which further reduces the demand and revenue. Second, the significant costs of clinical trial development constitute 94% of megafund investment and limit its financial performance. One potential solution is to use more cost-effective clinical trial designs such as adaptive trials [67] and platform trials [68], which simultaneously test multiple vaccine candidates using a shared control arm. These designs have been shown to significantly reduce clinical trial costs and expedite the drug development process [26]. Additionally, these novel trial designs do not share the ethical burden of human challenge trials.

We also note that the primary goal of the vaccine megafund is to prevent future infectious disease outbreaks and minimize the loss to global society. In light of this, we invest in clinical trials for all vaccine candidates simultaneously without optimizing the financial performance using sophisticated investment strategies [58] or financial engineering techniques such as dynamic leverage [69]. For example, if three vaccine candidates for the same infectious disease successfully complete their phase 2 trials, we may first conduct phase 3 trials for two vaccine candidates, initiating the phase 3 trial for the third vaccine only if the first two fail. This will reduce the costs of late-stage clinical trial development and improve its financial value. However, the increased financial value must be weighed against potential delays in FDA approvals of life-saving vaccines. A robust and multi-criteria optimization framework is needed to ensure that the value to society is not compromised by optimizing financial returns for the investors.

3.6 Conclusion

Despite the increased probability of success due to mRNA technology, diversification across a large number of vaccine candidates, and the potential benefits of conducting human challenge trials, the vaccine megafund model does not generate positive financial value for private-sector investors. Two limitations of the financial performance are the limited revenue of vaccine sales and the significant costs of late-stage clinical trial development. Nonetheless, the vaccine megafund generates significant societal value by preventing future epidemic outbreaks; if endowed with public sector funding of $10 billion, it may also generate positive financial value for investors.

These results underscore the urgency for continued collaboration between government agencies and the private sector in creating a sustainable business model and global vaccine ecosystem for addressing future pandemics. Stockpiling vaccines for the most dangerous EIDs, putting in place advance market commitments to purchase mass quantities of vaccines in case of outbreaks, creating government-sponsored manufacturing and distribution facilities that can supplement private-sector resources, and providing limited government guarantees to investors funding vaccine programs for a pre-specified list of priority diseases may all play a role in helping us reduce the impact of, or event prevent, future pandemics.

Part III

Statistical Innovations for Clinical Trial Design

Chapter 4

Bayesian Adaptive Clinical Trials for Anti-Infective Therapeutics during Epidemic Outbreaks

In the midst of epidemics such as COVID-19, therapeutic candidates are unlikely to be able to complete the usual multi-year clinical trial and regulatory approval process within the course of an outbreak. We apply a Bayesian adaptive patient-centered model—which minimizes the expected harm of false positives and false negatives—to optimize the clinical trial development path during such outbreaks. When the epidemic is more infectious and fatal, the Bayesian-optimal sample size in the clinical trial is lower and the optimal statistical significance level is higher. For COVID-19 (assuming a static $R_0 = 2$ and initial infection percentage of 0.1%), the optimal significance level is 7.1% for a clinical trial of a non-vaccine anti-infective therapeutic clinical trial and 13.6% for that of a vaccine. For a dynamic R_0 decreasing from 3 to 1.5, the corresponding values are 14.4% and 26.4%, respectively. Our results illustrate the importance of adapting the clinical trial design and the regulatory approval process to the specific parameters and stage of the epidemic.[1]

[1]Joint work with Shomesh E. Chaudhuri, Danying Xiao, and Andrew W. Lo. This chapter was published in *Harvard Data Science Review* [27]. We thank Murray Sheldon, Chi Heem Wong, the editor, associate editor, copyeditor, and several reviewers for many helpful comments and suggestions, and Jayna Cummings and Steven Finch for editorial assistance.

4.1 Introduction

With growing public concern over the outbreak of Coronavirus Disease 2019 (COVID-19), significant efforts have been undertaken by global biomedical stakeholders to develop effective diagnostics, vaccines, anti-viral drugs, medical devices, and other therapeutics against this highly infectious and deadly pandemic. While in the past, the traditional randomized clinical trial (RCT) and regulatory approval process often took several years [70]—longer than the typical duration of an epidemic outbreak [71]—recently the FDA has responded with actions such as the Breakthrough Devices Program, Emergency Use Authorization (EUA) authority, and Immediately in Effect guidance documents to prevent novel diagnostics and therapeutics from lagging behind the urgent needs of the population. In this paper, we propose adapting yet another tool that the FDA has already been exploring for medical devices [72], [73] to therapeutics for treating COVID-19 that are currently under development.

In recent years, Bayesian adaptive RCT protocols have been increasingly used to expedite the clinical trial process of potentially transformative therapies for diseases with high mortality rates [74]. Currently, these protocols have mainly been applied within the oncology domain, such as I-SPY for breast cancer [75] and GBM AGILE for glioblastoma [54]. These studies use Bayesian inference algorithms to greatly reduce the number of patients needed to assess the therapeutic effects of a drug candidate, without lowering the statistical power of the final approval decision, as measured by Type I and II error rates. As a result, therapeutic candidates can progress more quickly through the regulatory process and reach patients faster and at lower costs.

For severe diseases with no curative treatments, such as pancreatic cancer, patients tend to tolerate a higher Type I error of accepting an ineffective therapy in exchange for a lower Type II error of rejecting an effective therapy as well as expedited approvals of potentially effective treatments. Based on this observation, a patient-centered Bayesian protocol was proposed [14], [15] that incorporates patient values into clinical trial design and identify the optimal balance between the possibilities of false positives (Type I error) and false negatives (Type II error). For

more severe diseases, this protocol sets a tolerated Type I error rate much larger than the traditional 5% threshold, which leads to higher rates of approvals and expedited approval decisions.

However, the original Bayesian adaptive RCT framework does not take into account patient risk preferences. To address this gap, Chaudhuri and Lo [73] developed an adaptive version of the Bayesian patient-centered model that achieves an optimal balance between Type I and Type II error rates, significantly reducing the number of subjects needed in trials to achieve a statistically significant conclusion. A key feature of this model is the time evolution of the loss function of the Bayesian decision algorithm. This mechanism favors the expedited approval of diagnostic or therapeutic candidates that show early positive effects, since patients place a lower value on delayed approval of an effective diagnostic or therapy.

There is a natural but subtle analog to this dilemma in the case of therapeutics for an infectious disease during the course of an epidemic outbreak. Approving an effective therapeutic early will prevent future infections and deaths, while approving it later will save fewer people from infection. On the other hand, approving an ineffective therapeutic early will not prevent any future casualties. Worse still, it may prevent people from taking adequate precautions against infection, since they will falsely believe that they are safe from the disease after the advent of the ineffective therapy.

Moreover, the cost of Type I versus Type II error can differ from therapy to therapy. A novel vaccine that could trigger a significant immune response such as a cytokine storm has a much higher cost of a Type I error than a medical device such as an air filtration system designed to destroy virions through intense ultraviolet light. Therefore, the appropriate statistical threshold for approval should depend on the specific therapy, as well as the circumstances of the current burden of disease.

We apply the Bayesian adaptive protocol to anti-infective therapeutic development using a loss function that evolves over the course of an epidemic outbreak. We achieve an optimal balance between Type I and Type II errors for therapeutics that treat infectious diseases and identify the optimal time to reach the approval decision

based on the accumulation of clinical evidence. Our results show that when the epidemic is more infectious, the necessary sample size of the RCT decreases, while the tolerable Type I error increases. This confirms our earlier intuition that potentially effective therapies that are known to be safe should receive expedited approval when an epidemic is spreading rapidly.

4.2 Multi-Group SEIR Epidemic Model

The starting point for our analysis is the Susceptible-Exposure-Infective-Removed (SEIR) epidemic model, which has been applied to model the outbreak of COVID-19 in China in a number of recent studies [76], [77]. The population of N subjects is partitioned into four distinct groups: susceptible (S), exposed (E), infectious (I), and removed (R). The time evolution of the epidemic is specified by the following group of ordinary differential equations:

$$\frac{dS}{dt} = -\beta SI, \quad \frac{dE}{dt} = \beta SI - aE, \quad \frac{dI}{dt} = aE - \gamma I, \quad \frac{dR}{dt} = \gamma I \tag{4.1}$$

Here we use the convention that $S(t), E(t), I(t)$, and $R(t)$ are the *proportions* of the susceptible, exposed, infectious and removed populations, respectively, satisfying the conservation constraint for all t:

$$S(t) + E(t) + I(t) + R(t) = 100\% \tag{4.2}$$

The parameters β, a, and γ denote the average rates of infection, incubation and recovery, respectively, and $\mu \in (0\%, 100\%)$ denotes the mortality rate of the epidemic. For example, if $\mu = 5\%$, we expect 5% of infected subjects will die from the disease. At time t, $\mu R(t)N$ subjects will have died, and $(1-\mu)R(t)N$ will have recovered.

A critical measure of the infectivity of an epidemic is its basic reproduction number, defined as $R_0 = \beta/\gamma$ in the SEIR model. This is the expected number of secondary infections caused by each infected subject in a population with no public health measures (such as quarantine, social-distancing, or vaccination).

A number of studies have used different statistical schemes to estimateR_0 for COVID-19 during its initial outbreak period in central China in January 2020. These estimated values of R_0 range from 2.2 (95% CI, 1.4 to 3.9) [78] to 3.58 (95% CI, 2.89 to 4.39) [79]. Given the large uncertainty in the value of R_0, we simulate therapeutic development under scenarios with constant R_0 values of 2 and 4.

In addition, to model the impact of governmental nonpharmaceutical interventions (NPIs) on containing the spread of the epidemic, we consider a dynamic transmission SEIR model where the infection rate $\beta(t)$ monotonically decreases in time as a result of the NPI. Specifically, we assume that $\beta(t)$ takes the sigmoid functional form:

$$\beta(t) = \frac{\beta_0 - \beta_\infty}{1 + \exp\left(\frac{t-t_2}{\tau}\right)} + \beta_\infty \tag{4.3}$$

Here β_0 and β_∞ denote the infection rates in the initial and final stages of the epidemic (with $\beta_0 > \beta_\infty$), respectively, t_2 denotes the half-life of the decay in infection rate, and τ the length of the time window when this decay occurred. A larger difference $\beta_0 - \beta_\infty$ corresponds to more significant reduction of epidemic transmission, a smaller value of t_2 corresponds to a speedier decision to enforce the NPI, and a smaller value of τ corresponds to more strict enforcement of the NPI since $\beta(t)$ decays more rapidly. We calibrate the values $\beta_0 = 3$ and $\beta_\infty = 1.5$ based on the estimates of the dynamic transmission rate of COVID-19 in Wuhan, China from December 2019 to February 2020 [80]. We consider different values of t_2 and τ to reflect the variability in timing and stringency of NPIs enforced by governments around the globe. Under this dynamic transmission model, the basic reproduction number is given by $R_0(t) = \beta(t)/\gamma$, which monotonically decreases from β_0/γ to a constant value β_∞/γ as t increases.

To model the significant variability in mortality rates of COVID-19 for patients in different age groups, we extend this basic model to a multi-group SEIR model, where the population is partitioned into five age groups, (1) below 49, (2) 50 to 59, (3) 60 to 69, (4) 70 to 79, and (5) above 80. We use S_i, E_i, I_i, and R_i to denote the corresponding type in each group (and continue to use S, E, I, and R for the total

proportion of each type in all groups). The dynamics of the epidemic are specified by the modified ordinary differential equations:

$$\frac{dS_i}{dt} = -\beta c_i S_i I \quad \frac{dE_i}{dt} = \beta c_i S_i I - aE_i \quad \frac{dI_i}{dt} = aE_i - \gamma I_i \quad \frac{dR_i}{dt} = \gamma I_i \tag{4.4}$$

Here c_i denotes the contact rate of the susceptible subjects in the i^{th} age group with the total infected population I of all groups. This contact rate is measured relative to group 1, which we normalize to $c_1 = 1$. In the case of COVID-19, although the mortality rate is much higher for senior populations [81], the elderly also tends to have less frequent contact with the infected population outside the household [82]. We solve the differential equations in the multi-group SEIR model using the ODE45 solver in MATLAB 2019a with initial conditions for each age group:

$$[S_i(0),\ E_i(0)\ , I_i(0),\ R_i(0)] = [1 - (1 + r_e)I_0, r_e I_0, I_0, 0] \times P_i \tag{4.5}$$

The parameter I_0 denotes the proportion of the initially infected population, r_e is the ratio of initially exposed and infected subjects, and P_i is the percentage of the i^{th} age group in the population. The assumed demographic, contact rate, and mortality rate values are summarized in Table 4.2.

Table 4.1: Demographic profile of various age groups for COVID-19, SARS, and MERS in the U.S. population.

Age Group	Percentage (%)	Contact Rate (c_i)	Disease	Mortality (%)
Below 49	64	1.00	COVID-19	0.3
			SARS	3
			MERS	15
50 to 59	13	0.83	COVID-19	1.3
			SARS	10
			MERS	30
60 to 69	12	0.66	COVID-19	3.6
			SARS	17.6
			MERS	35
70 to 79	7	0.50	COVID-19	8
			SARS	28
			MERS	45
Above 80	4	0.42	COVID-19	14.8
			SARS	26.3
			MERS	40

Table 4.2: Simulation parameters and values.

Parameter	Description	Value(s)
R_0	Basic reproduction number	$2, 4$
a	Incubation rate (per week)	1
γ	Recovery rate (per week)	1
I_0	Initial proportion of infected population	$0.1\%, 0.01\%$
r_e	Ratio of initially exposed and infected populations	10
$[\beta_0, \beta_\infty]$	Initial and final infection rate in dynamic model	$[3, 1.5]$
t_2	Half-life of decay in the dynamic model (week)	$3, 6$
τ	Window length of decay in the dynamic model (week)	$0.5, 1$
N	Population size (million)	300

4.3 A Bayesian Patient-Centered Approval Process

Similar to Chaudhuri and Lo [73], we develop a Bayesian patient-centered decision model for RCT approval which minimizes the expected loss (or harm) incurred on the patients by optimally balancing the losses of Type I and Type II errors. Here the loss does not refer to financial costs afforded by the patients, but rather the loss in patient value (i.e. how much patients weigh the relative harms of infection and death). We assign the losses per patient of being susceptible, infected, and deceased. Since Bayesian decision thresholds are invariant under the rescaling of the losses, we normalize by setting the loss per patient infection $L_I = 1$. We then assign the loss per patient death relative to L_I as L_D, and the loss due to susceptibility to the disease as L_S. The parameter values we assume, summarized in Table 4.1, are meant to represent one reasonable valuation of the relative losses. However, in practice patient value will differ from one patient group another, especially given the large variability of mortality rate of COVID-19 in different age groups [81]. Here we report the main results of optimal sample size and statistical significance (Table 4.4 and 4.5) assuming $L_D = 100$. The results for $L_D = 10$ are provided in Tables B.3 and B.4.

We simulate the multi-group SEIR model over a time period of T weeks, where T

is the duration of the epidemic outbreak. Let κ denote the weekly subject enrollment rate in each arm of the clinical trial. We assume that the value of R_0 is known (or well-estimated) at initial time t=0 and stays constant during the course of the outbreak. At time $t \in [0,\ T]$, the Bayesian loss $C_{ij}(t)$ of choosing the action $\hat{H} = i$ under $H = j$ is defined as:

Table 4.3: Cost matrix of Bayesian decision analysis.

	$\hat{H} = 0$ (do not approve)	$\hat{H} = 1$ (approve)
$H = 0$ (no effect)	0	$(S(t) - S(T))NL_S$
$H = 1$ (effective)	$R(T)N(L_I + \mu L_D)$	$CI(t)NL_I + \mu NR(t)L_D$

where we define the cumulative number of infected patients $CI\,(t)$ until time t:

$$CI\,(t) = E\,(t) + I\,(t) + R\,(t) \tag{4.6}$$

By design, this loss function penalizes Type I errors early in the epidemic by the susceptible term, $(S\,(t) - S(T))NL_S$. We subtract the base level $S(T)$ from $S\,(t)$ since the multi-group SEIR model predicts that $S(T)N$ subjects will not be infected by the epidemic. A Type I error at an earlier time will expose more currently susceptible population to the epidemic, since they will falsely believe that they are safe from the disease after the advent of the ineffective therapeutic. On the other hand, the loss function also penalizes correct approval decisions made at later stages of an epidemic via to the cumulative infected and death terms, $CI\,(t)$ and $\mu R(t)$. A correct but delayed approval decision for the therapeutic is less valuable since it will save fewer susceptible people from infection and death.

The Bayesian decision model considers the null hypothesis $H = 0$ that the anti-infective therapeutic (or vaccine) has no clinical effect, against the alternative hypothesis that it has positive clinical effect with signal-to-noise ratio ρ [73]. We use p_0 and p_1 to denote the Bayesian prior probabilities of $H = 0$ and $H = 1$, respectively.

This patient-value model imposes higher losses for incorrect approvals at earlier stages and correct approvals at later stages of an epidemic. Under these constraints, the Bayesian decision algorithm yields the sample size and statistical significance

threshold of the RCT that optimally balances Type I and Type II error.

4.4 Results

We simulate an epidemic outbreak over a time period of T weeks, where T is the duration of the outbreak. For an epidemic with higher infectivity, its duration is shorter, which puts more pressure to reach a timely approval decision. To avoid numerical instability, we formally define T as the first time when the number of cumulative infected subjects reaches 99.9% of total infections predicted by the SEIR model. We assume an age-specific mortality rate μ at the level of COVID-19 [81], [83], and incubation and recovery periods of 7 days each [76]. These estimated parameters can all be challenged to varying degrees, depending on the specific drug-indication pair under consideration and the particular circumstances of the epidemic, but they are meant to be representative for a typical anti-infective therapeutic during the midst of a growing epidemic.

We also assume that it takes 7 days after injection to assess the efficacy of the therapeutic on each subject. We adopt the optimization scheme of [15] to find the optimal Type I and Type II error rates of the non-adaptive Bayesian RCT. To represent typical practice of the pharmaceutical industry, we optimize under the upper bound on the model's power $\text{Power}_{\max} = 90\%$ [14]. We then use these optimal error rates as our stopping criteria to simulate the sequential decision process of a Bayesian adaptive RCT via Monte Carlo simulation [73]. The simulation results are summarized in Tables 4.4 and 4.5.

We separate the results into two distinct types of therapeutics—non-vaccine anti-infectives (Table 4.4) and vaccines (Table 4.5)—because of the differences in their historical probabilities of success. Vaccine development programs have an estimated probability of success $p_1^{vac} = 40\%$ as of 2019Q4 [3] whereas the corresponding figure for non-vaccine anti-infectives is $p_1^{nv} = 23\%$ [23].

Table 4.4: Simulation results of a Bayesian adaptive RCT on non-vaccine anti-infective therapeutics.

Epidemic Parameters				Nonadaptive		Adaptive sample size H_0		Adaptive sample size H_1			
R_0	μ	I_0 (%)	n^*	α^* (%)	Power (%)	Mean (SD)	Median (IQR)	Mean (SD)	Median (IQR)	α (%)	Power (%)
2	COVID-19	0.1	242	7.1	90	135 (103)	105 (63, 176)	148 (107)	119 (73,192)	5.8	91.5
4	COVID-19	0.1	158	17.3	90	115 (83)	91 (56,149)	98 (80)	74 (42,128)	14.4	92.1
$R_0(t)$	COVID-19	0.1	176	14.4	90	118 (85)	95 (57, 153)	108 (85)	84 (49,140)	11.7	92.2
2	COVID-19	0.01	399	1.2	90	150 (128)	110 (64, 191)	248 (154)	211 (139,317)	1.0	91.0
4	COVID-19	0.01	274	5.0	90	140 (110)	106 (64, 180)	168 (116)	136 (86,216)	4.1	91.4
$R_0(t)$	COVID-19	0.01	304	3.6	90	145 (119)	108 (64, 187)	184 (120)	153 (97,239)	3.0	91.2
2	SARS	0.1	164	16.3	90	117 (85)	94 (57, 150)	101 (81)	77 (45,132)	13.9	92.3
4	SARS	0.1	112	27.8	90	98 (72)	79 (47, 128)	72 (64)	51 (27,95)	23.3	93.2
$R_0(t)$	SARS	0.1	107	29.2	90	96 (71)	78 (45, 126)	70 (64)	50 (26,92)	25.1	93.4
2	MERS	0.1	88	35.3	90	87 (66)	70 (40, 115)	59 (57)	39 (20,77)	29.8	93.7
4	MERS	0.1	63	45.2	90	73 (59)	59 (30, 100)	46 (49)	28 (14,60)	38.8	94.5
$R_0(t)$	MERS	0.1	44	54.3	90	61 (54)	48 (20, 86)	36 (43)	19 (9,46)	47.0	94.8

Results are obtained from 10,000 Monte Carlo runs and assuming $L_D = 100$. R_0 denotes the basic reproduction number, μ the disease morality, and I_0 the proportion of initial infected subjects. Sample size refers to the number of subjects enrolled in each arm of the RCT. SD denotes standard deviation, and IQR the interquartile range about the median. $R_0(t)$ denotes the dynamic transmission model with $t_2 = 3$ weeks, $\tau = 1$ week, and $R_0(t)$ decreasing from 3 to 1.5 as time t increases.

Table 4.5: Simulation results of a Bayesian adaptive RCT on vaccines.

Epidemic Parameters				Nonadaptive		Adaptive sample size H_0		Adaptive sample size H_1			
R_0	μ	I_0 (%)	n^*	α^* (%)	Power (%)	Mean (SD)	Median (IQR)	Mean (SD)	Median (IQR)	α (%)	Power (%)
2	COVID-19	0.1	181	13.6	90	122 (91)	95 (58, 158)	112 (87)	86 (51,145)	11.3	92.4
4	COVID-19	0.1	111	28.1	90	97 (71)	78 (47,127)	72 (64)	52 (27,95)	23.4	92.8
$R_0(t)$	COVID-19	0.1	117	26.4	90	71 (49)	80 (49, 132)	74 (67)	53 (29,98)	21.8	93.1
2	COVID-19	0.01	342	2.3	90	148 (124)	110 (64, 191)	212 (137)	177 (115,275)	2.1	90.9
4	COVID-19	0.01	232	7.9	90	132 (100)	104 (61, 171)	142 (103)	113 (69,184)	6.6	91.4
$R_0(t)$	COVID-19	0.01	244	6.9	90	132 (101)	102 (63, 171)	148 (106)	118 (73,191)	5.4	91.7
2	SARS	0.1	99	31.7	90	91 (67)	74 (44, 119)	65 (62)	44 (24,86)	26.6	93.5
4	SARS	0.1	65	44.3	90	74 (59)	60 (32, 99)	47 (50)	29 (14,61)	37.6	94.5
$R_0(t)$	SARS	0.1	50	51.3	90	65 (55)	52 (23, 91)	40 (45)	23 (11,51)	44.2	94.8
2	MERS	0.1	27	64.2	90	49 (49)	36 (11, 71)	28 (37)	13 (6,34)	58.9	95.9
4	MERS	0.1	21	68.1	90	45 (48)	31 (8, 66)	25 (35)	11 (5,29)	58.9	96.3
$R_0(t)$	MERS	0.1	7	79.2	90	33 (41)	14 (4, 50)	17 (27)	6 (3,18)	69.1	97.2

Results are obtained from 10,000 Monte Carlo runs and assuming $L_D = 100$. R_0 denotes the basic reproduction number, μ the disease morality, and I_0 the proportion of initial infected subjects. Sample size refers to the number of subjects enrolled in each arm of the RCT. SD denotes standard deviation, and IQR the interquartile range about the median. $R_0(t)$ denotes the dynamic transmission model with $t_2 = 3$ weeks, $\tau = 1$ week, and $R_0(t)$ decreasing from 3 to 1.5 as time t increases.

4.4.1 Non-Vaccine Anti-Infective Therapeutics

Static Transmission Rate

We first analyze the case when the infectivity R_0 remains constant in time (e.g. in the absence of effective NPIs). For the fixed-sample Bayesian RCT on a non-vaccine anti-infective therapeutic, as R_0 increases from 2 to 4 (Rows 1 to 2 of Table 4.4), the optimal sample size of each experimental arm decreases from 242 to 158 and the optimal Type I error rate drastically increases from 7.1% to 17.3% (Figure 4-1), much higher than the traditional 5% threshold. As the epidemic spreads across the population more rapidly, the Bayesian RCT model has greater pressure to expedite the approval process and a much higher tolerance of false positive outcomes.

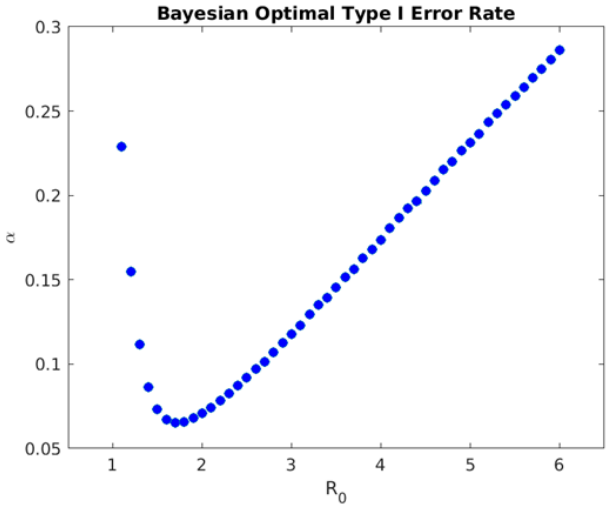


Figure 4-1: Optimal Type I error rate α of a non-adaptive Bayesian RCT vs. basic reproduction number R_0.

For the Bayesian adaptive RCT, when the therapeutic is ineffective ($H = 0$), the average sample size required to reject the therapeutic is much smaller than that of the non-adaptive version (Columns 7 and 8 of Table 4.4). Also, the required sample size decreases with the infectivity R_0 in both mean and quartiles, yet always achieves Type I error rate α below that of the non-adaptive version (Column 11). The adaptive Bayesian decision model is able to reject an ineffective therapeutic with a relatively small sample size and a bounded false-positive rate.

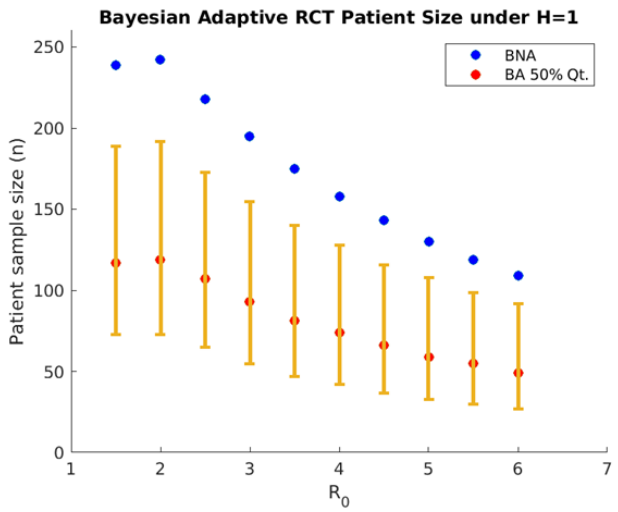


Figure 4-2: Subject sample size in each arm of a Bayesian adaptive RCT under $H = 1$ decreases with the basic reproduction number R_0.

On the other hand, when the therapeutic is effective ($H = 1$), as R_0 increases from 2 to 4, the average sample size required by the Bayesian adaptive RCT decreases from 148 to 98 (Columns 9 and 10 of Table 4.4). The Bayesian adaptive model places more weight on approving an effective therapeutic earlier to prevent future infections when the epidemic is more infectious. Despite the smaller sample size, the model still retains an empirical power above 91.0% for all values of R_0 (Column 12). The Bayesian adaptive model simultaneously expedites the approval of an effective therapeutic and retains a bounded false-negative rate. The results are illustrated in Figure 4-2.

Furthermore, as the proportion of the initially infected population I_0 decreases from 0.1% to 0.01% (Rows 4 to 6 of Table 4.4), the optimal sample sizes for non-adaptive and adaptive RCTs both increase, while the optimal Type I error rates decrease. Beginning the clinical trials for a therapeutic during the earlier stages of an epidemic outbreak reduces the need to expedite the approval process in order to contain its future spread. Clinicians and researchers have more time to evaluate the efficacy of a therapeutic and record adverse effects by testing it on a larger number of subjects, which leads to a lower Type I error rate.

Finally, when the mortality rate μ increases from the level of COVID-19 [81], [83],

to the level of SARS [84], and further to the level MERS [85], the optimal sample sizes for both non-adaptive and adaptive Bayesian models decrease and the optimal Type I error rates increase (Rows 7 to 12 of Table 4.4). When the epidemic is more lethal, the Bayesian adaptive model requires fewer subjects in the RCT, since both Type I and Type II errors will lead to greater losses due to death by infection. The higher death tolls provide significantly more incentive in the Bayesian adaptive framework to approve the therapeutic in the hopes of saving more people from future infection and death.

One interesting feature of the Bayesian decision model is that the optimal Type I error rate is not a monotonic function R_0, but rather has a global minimum of 8% at $R_0 = 1.7$ for COVID-19, as shown in Figure 4-1. As R_0 decreases below 1.7, the optimal Type I error rate increases. The intuition for this result is that we define the loss of Type I error as the excess risk of being susceptible to infection $(S(t) - S(T))NL_S$, where $S(T)$ is the fraction of the population that remains uninfected throughout the epidemic outbreak. When R_0 is small, $S(T)$ is close to 100% and the excess risk $(S(t) - S(T))NL_S$ is small compared to the benefit of preventing future deaths. Therefore, when the epidemic is not very infectious, the Bayesian decision model expedites the approval decision. This also confirms the intuition that smaller sample sizes are required in adaptive trials for diseases that affect a small fraction of the population. If we instead define the loss of Type I error as the absolute risk of being susceptible $S(t)NL_S$, we find that the optimal Type I error indeed monotonically increases with R_0, as shown in Figure B-3.

Dynamic Transmission Rate

The results for the dynamic transmission model with $\beta_0 = 3, \beta_\infty = 1.5, t_2 = 3$ weeks and $\tau = 1$ week are summarized in Table 4.4. For COVID-19 (rows 3 and 6 in Table 4.4), we find that the Bayesian optimal sample size and Type I error rate of the dynamic transmission model lie in between the corresponding values under scenarios $R_0 = 2$ and $R_0 = 4$. This suggests that timely and effective government interventions will protect more subjects from infection and allow more time for the RCT.

However, for the more fatal SARS and MERS (rows 9 and 12 in Table 4.4), the dynamic transmission model sets higher optimal Type I error α and smaller sample size than $R_0 = 4$. This is due to the U-shaped curve of optimal α vs. R_0, shown in Figure 4-1. When the NPI reduces $R_0(t)$ below a certain threshold, the optimal α starts to increase. For highly fatal epidemics, when the government adopts NPIs to protect most of the susceptible population from infection, the regulatory priority should be to expedite potentially effective treatments that can help current patients since the loss of Type I error is much lower than that of the Type II error.

In addition, we investigate the impact of the timing and stringency of NPIs enforced by the government with different values of t_2 and τ. The results are summarized in Table 4.6. We find that the optimal Type I error is larger for $t_2 = 3$ weeks than $t_2 = 6$ weeks. Therefore, if the government adopts well-enforced NPIs early on (such as the lockdown of Wuhan, China) to protect the susceptible population, this will reduce the loss associated with Type I error, leading to expedited approvals of potentially effective therapeutics. Furthermore, the sooner an effective therapeutic is approved, the sooner will NPIs be lifted.

Table 4.6: Optimal sample size and Type I error α of Bayesian non-adaptive RCT for non-vaccine anti-infective therapeutics for dynamic transmission model.

Disease[2]	$I_0(\%)$	R_0	t_2 (week)	τ (week)	Sample Size	α^* (%)	Power (%)
COVID-19	0.1	2	NA	NA	242	7.1	90
		4	NA	NA	158	17.3	90
		$R_0(t)$	3	0.5	166	16.0	90
		$R_0(t)$	3	1	176	14.4	90
		$R_0(t)$	6	0.5	176	14.4	90
		$R_0(t)$	6	1	177	14.2	90
SARS	0.1	2	NA	NA	164	16.3	90
		4	NA	NA	112	27.8	90
		$R_0(t)$	3	0.5	100	31.3	90
		$R_0(t)$	3	1	107	29.2	90
		$R_0(t)$	6	0.5	118	26.2	90
		$R_0(t)$	6	1	119	25.9	90
MERS	0.1	2	NA	NA	88	35.3	90
		4	NA	NA	63	45.2	90
		$R_0(t)$	3	0.5	41	55.9	90
		$R_0(t)$	3	1	44	54.3	90
		$R_0(t)$	6	0.5	59	47.0	90
		$R_0(t)$	6	1	60	46.5	90

4.4.2 Vaccines

We repeat the above analysis for RCT of vaccines using a prior probability of having an effective vaccine $p_1^{vac} = 40\%$ as reported by Project ALPHA in 2019Q4 [3]. The simulation results are summarized in Table 4.5. Overall, we observe the same pattern in the optimal sample size and Type I error rates on infectivity, mortality, and

[2] R_0 denotes the basic reproduction number, μ the disease morality, and I_0 the proportion of initial infected subjects. Sample size refers to the number of subjects enrolled in each arm of the RCT. $R_0(t)$ denotes the use of a dynamic transmission model with $\beta_0 = 3$ and $\beta_\infty = 1.5$.

proportion of initial infections. However, since p_1 is higher for vaccines, the Bayesian decision model requires fewer subjects on average in the RCT to ascertain the positive effects of the vaccine, compared to the case of anti-infective therapeutics in Table 4.4. We find that vaccines should receive even more expedited evaluation.

4.4.3 Five-Factor Sensitivity Analysis

To assess the robustness of our model's predictions against the assumed values of model parameters, we perform a five-factor sensitivity analysis for the static transmission rate model with $R_0 = 2$. The baseline and alternative parameter values are summarized in Table B.1. The scatter plot of optimal Type I error α vs. sample size of Bayesian non-adaptive RCT model is shown in Figure 4-3. We find that the scatter plot consists of several curves. To clearly identify the effect of any given parameter, we show the results for the most important parameters in separate scatter plots in Figures 4-4, 4-5, and 4-6.

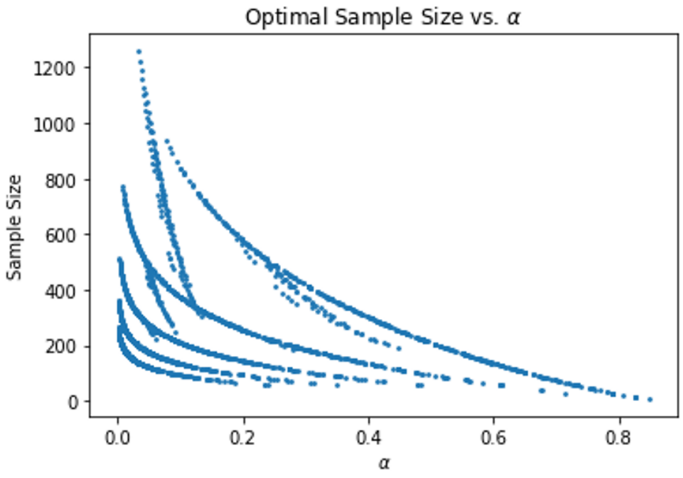


Figure 4-3: Scatter plot and summary statistics of optimal Type I error α vs. optimal sample size from the five-factor analysis when $R_0 = 2$.

We find that the different curves in Figure 4-3 result from different values of ρ, the signal-to-noise ratio (SNR) of treatment effect [73], as shown in Figure 4-4. For a given significance level α, a smaller value of ρ leads to larger optimal sample size. If the efficacy of the anti-infective therapeutic is insignificant (small ρ), the distributions of

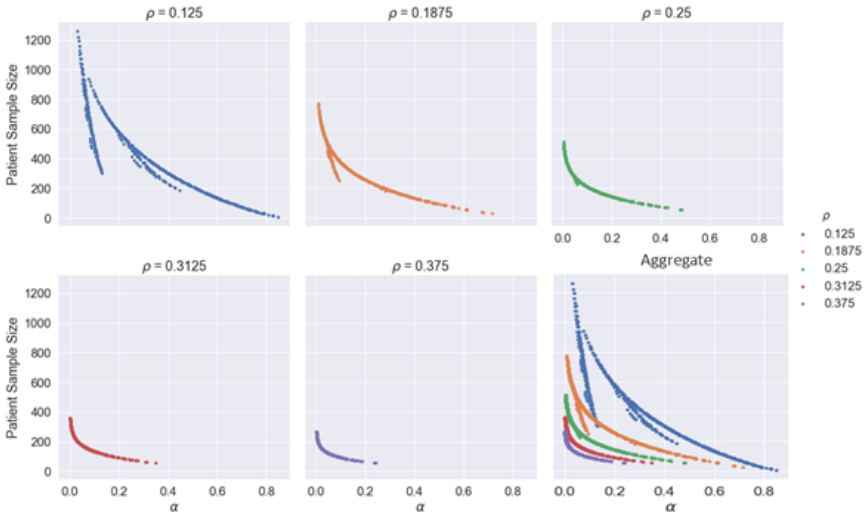


Figure 4-4: Scatter plot of optimal Type I error rate α vs. sample size for different values of ρ, signal-to-noise ratio of the treatment effect [72].

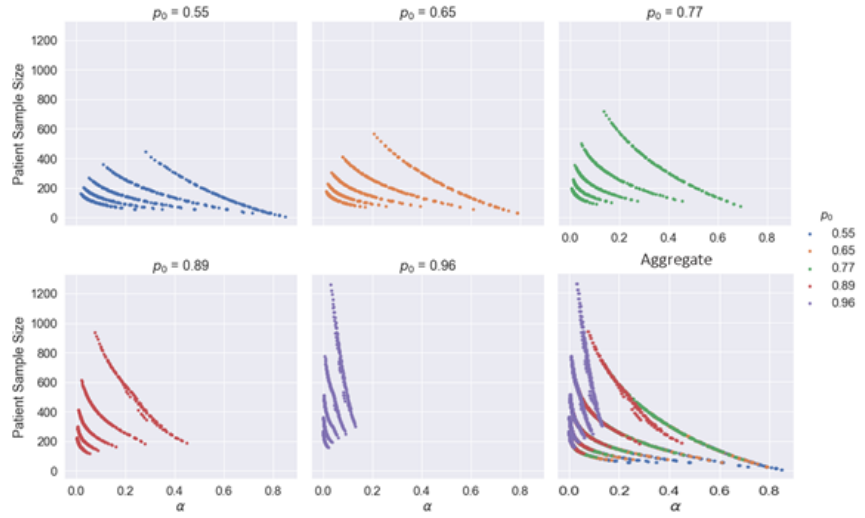


Figure 4-5: Scatter plot of optimal Type I error rate α vs. sample size for different values of p_0, Bayesian prior probability of having an ineffective therapeutic.

z-score under the null hypothesis $H = 0$ (no effect) and alternative hypothesis $H = 1$ (positive effect with SNR ρ) are difficult to distinguish statistically. Hence a larger sample size is needed to evaluate the efficacy at the given significance level α.

In addition, with a fixed SNR ρ, the magnitudes of α and sample size are mainly determined by p_0^{nv}, the Bayesian prior probability of having an ineffective anti-infective therapeutic. A larger value of p_0^{nv} leads to a smaller α and a larger sample size (Figure

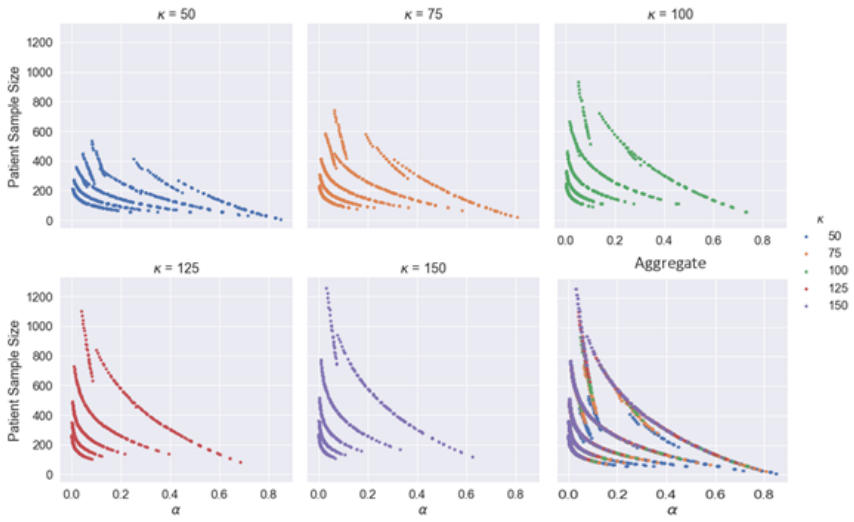


Figure 4-6: Scatter plot of optimal Type I error rate α vs. sample size for different values of κ, weekly patient enrollment rate (patients per week) in each arm of RCT.

4-5). When past drug development outcomes in the anti-infective domain strongly suggest that the current anti-infective therapeutic is unlikely to be effective (large p_0^{nv}), the Bayesian framework requires many more observations to shift the posterior distribution in order to prove its efficacy.

A similar but less significant effect on the magnitudes of α and sample size is generated by κ, the weekly subject enrollment in each arm of RCT. A larger value of κ leads to a smaller α and a larger sample size (Figure 4-6). When the RCT enrolls patients at a faster rate, clinical researchers may evaluate the efficacy of the treatment based on more observations early on during the epidemic outbreak. Hence the approval decision may be reached with lower false positive error.

In Figures B-2 and B-1, we show that the five-factor sensitivity analysis reveals no significant dependence of α and sample size on Δt, the time needed to assess the treatment efficacy, as well as a, the incubation period of the disease.

However, the analysis does show that extreme values of the Bayesian optimal Type I error rate are generated by large values of p_0^{nv} and small values of ρ. These regions of parameter space can be avoided if the anti-infective therapeutic under investigation has promising preclinical evidence to support its efficacy (reducing p_0^{nv}) and the RCT is designed to verify reasonably significant treatment effects over the control arm

(increasing ρ).

4.5 Discussion

A natural consequence of using a patient-centered framework for determining the approval threshold is, of course, more false positives—and the potential for a greater number of patients with adverse side-effects—in cases where the burden of disease is high. These false positives can be addressed through more vigilant post-approval surveillance by regulatory agencies and greater requirements for drug and device companies to provide such patient-level data to the regulator following approval. Failure to provide such data or evidence of an ineffective therapy can be grounds for revoking the approval.

However, past experience shows that withdrawing an approved drug can be challenging and disruptive for several reasons [86]. Therefore, implementing the patient-centered approach may require creating a new category of temporary approvals for crisis situations involving urgent needs at national or international levels, similar to the FDA's EUA program. Such a program might involve provisional approval of a candidate therapy consisting of a one- or two-year license—depending on the nature of the drug-indication pair—to market the therapy to a pre-specified patient population, no off-label use of the therapy, and regular monitoring and data reporting to the regulator by the manufacturer and/or patients' physicians during the licensing period [87]. At the end of this trial period, one of two outcomes would occur, depending on the accumulated data during this period: (a) the "urgent needs" license expires; or (b) the license converts to the traditional regulatory license. Of course, at any point during the trial period, the regulator can terminate the license if the data show that the therapeutic is ineffective and/or unsafe.

While such a process may impose greater burdens on patients, manufacturers, and regulators, it may still be worthwhile if it brings faster or greater relief to patients facing mortal illnesses and extreme suffering. In this respect, an urgent-needs program may be viewed as a middle ground between a standard clinical trial and

an approval, similar in spirit to the adaptive designs of sophisticated clinical trials with master protocols such as I-SPY 2, LUNG-MAP, and GBM-AGILE, in which patient care and clinical investigations are simultaneously accomplished. Also, because the Centers for Medicare and Medicaid Services (CMS) has demonstrated a willingness to cover the cost of certain therapeutics for which evidence is still being generated (see, for example, CMS's "coverage with evidence" programs listed at https://go.cms.gov/2v6ZxWm), additional economic incentives may be available to support such temporary licenses.

Finally, we note that the age-group specification in our SEIR model mainly focuses on older populations, whose mortality risks with COVID-19 are much higher than younger populations [81]. More refined age-group specifications are needed to differentiate the transmission rates of COVID-19 among children, teenagers, and young adults, as well as to reflect the different societal benefits each age group will receive from the approval of an effective anti-infective therapeutic or vaccine.

4.6 Conclusion

We apply the Bayesian adaptive patient-centered model of Chaudhuri and Lo [73] to clinical trials for therapeutics that treat infectious diseases during an epidemic outbreak. Using a simple epidemiological model, we find that the optimal sample size in the clinical trial decreases with the infectivity of the epidemic, measured by the basic reproduction number R_0. At the same time, the optimal Type I error rate increases with R_0. Lower levels of initial infection increase the number of subjects required to verify the therapeutic efficacy of the therapeutic under investigation, while higher levels of mortality increase the optimal sample size. The results confirm our intuition that clinical trials should be expedited and a higher false positive rate should be tolerated when the epidemic spreads more rapidly through the population, has a higher mortality rate, and has already infected a sizable portion of the population at the beginning of the RCT.

To provide transparency for how a patient-centered approach differs from the

traditional statistical framework in the anti-infectives context, we use a relatively simple mathematical model of epidemic disease dynamics to estimate the societal loss in an outbreak. More sophisticated epidemiological models can easily be incorporated into our framework at the cost of computational tractability and transparency.

One interesting trade-off to be explored is the difference between COVID-19 vaccine and an anti-viral treatment that can cure an infected patient. While prevention through vaccination is the ultimate solution, a successful treatment for the disease using repurposed drugs that have already been approved for other indications (and whose safety profile has already been established) may be even more valuable, especially if it can be deployed in the nearer term and reduce the growing fear and panic among the general population. In such cases, the approval threshold should clearly reflect these cost/benefit differences.

Of course, in practice regulators consider many factors beyond p-values in making its decisions. However, that process is opaque even to industry insiders, and the role of patient preferences is unclear. The proposed patient-centered approach provides a systematic, objective, adaptive, and repeatable framework for explicitly incorporating patient preferences and burden-of-disease data in the therapeutic approval process. This framework also fulfills two mandates for the FDA, one from the fifth authorization of the Prescription Drug User Fee Act (PDUFA) for an enhanced quantitative approach to the benefit-risk assessment of new drugs [88], and the other from Section 3002 of the 21st Century Cures Act of 2016 requiring the FDA to develop guidelines for patient-focused drug development, which includes collecting patient preference and experience data and explicitly incorporating this information in the drug approval process.

We hope this work will shed further insight into improving the current clinical trial process for infectious disease therapeutics and contribute to the timely development of effective treatments and vaccines for COVID-19 in particular.

Chapter 5

A Bayesian Decision Analysis of Phase 2 Clinical Trial Outcome of AMX0035 for Amyotrophic Lateral Sclerosis

Amyotrophic lateral sclerosis (ALS) is a motor neuron disease with no curative treatment and poor prognosis. Recently, the new drug application (NDA) of AMX0035 has generated significant controversy, since the drug has not completed its phase 3 trial, although its phase 2 trial has shown clear therapeutic effects (with p-value 3%) in slowing the rate of ALS progression. To determine whether the clinical evidence justifies an NDA for AMX0035, we apply Bayesian Decision Analysis (BDA) to determine the optimal tradeoff between false positives (Type I error) and false negatives (Type II error). We find that the BDA-optimal Type I error rate is higher than 3% when the prior probability of an effective drug is at least 30%. Assuming a 70% probability that the drug is effective and 33% signal-to-noise ratio of treatment effect, the optimal Type I error rate is 15.4%, four times higher than the observed value in the phase 2 trial. [1]

[1] Joint work with Andrew W. Lo.

5.1 Introduction

Amyotrophic lateral sclerosis (ALS) is a fatal motor neuron disease with no treatments to cure or reverse its disease progression. Typically, ALS progresses from muscle weakness to death by respiratory paralysis in 3 to 5 years [89]. In 2015, ALS affected more than 16,500 patients in the U.S. [90], and the nationwide economic burden of ALS is estimated to be over $1 billion [91]. In a statement on March 2, 2021, the U.S. Food and Drug Administration (FDA) stated that it is "prepared to use all expedited development and approval pathways available" to facilitate the development and approval of agents to treat ALS [92]. On November 2, 2021, the biotech company Amylyx submitted a New Drug Application (NDA) to the FDA for its investigational ALS therapeutic AMX0035 [37]. This NDA is controversial given that the drug candidate has not completed its ongoing phase 3 clinical trial [93], although its phase 2 trial, with a sample size of 137, has shown a clear and statistically significant reduction in the rate of progression of ALS [38], [39]. Encouraged by the phase 2 results and propelled by the urgency to treat more ALS patients with the drug, two ALS patient advocacy groups have submitted over 50,000 signatures to the FDA, calling on the agency to approve AMX0035 [94].

To determine whether the clinical evidence justifies the NDA for AMX0035, we apply Bayesian Decision Analysis (BDA) to compute the optimal tradeoff between a false positive and the potential side effects of an ineffective drug (Type I error) versus a false negative in which patients are not given access to an effective drug (Type II error). We find that the BDA-optimal Type I error most closely associated with the AMX0035 phase 2 trial parameters is 15.4%—which is considerably higher than the 3% p-value reported in that trial [38]—and even higher for other plausible measures of burden of disease.

5.2 Literature Review

The traditional approach to assessing the weight of statistical evidence of a randomized clinical trial is to compare the p-value of the standardized test statistic associated with the trial outcome against the desired false-positive or Type I error rate, usually 5% for a one-tailed hypothesis test and 2.5% for a two-tailed tested. Trial outcomes with p-values below this threshold are deemed statistically significantly different from the null hypothesis of no effect, leading to regulatory approval, whereas those above it are deemed statistically indistinguishable from the null and do not lead to approval.

The question raised and answered by BDA is "why 2.5% or 5%?" For fatal diseases with no existing treatments, patients may be willing to accept a much higher false positive rate, especially if it yields a lower false negative rate or Type II error, as is often the case. For example, suppose the conventional 5% Type I error is associated with a Type II error of 25%. A glioblastoma patient that has exhausted the standard of care and may be comfortable with a Type I error of 20% if it is associated with a Type II error of 10%. Given that such patients have no other recourse for this terminal illness, the relative importance of false positives and negatives should reflect their circumstances.

Regulatory authorities recognize the challenge facing desperate patients that have run out of options and have developed a number of mechanisms for expediting the approval process for drugs intended to treat the most serious conditions. For example, the FDA offers four clinical-trial designations—fast track, breakthrough therapy, accelerated approval, and priority review—that involve faster reviews and/or use surrogate endpoints to judge efficacy. However, published descriptions [95], [96] do not indicate any differences in the statistical thresholds used in these programs versus the standard approval process, nor do they mention adapting these thresholds to the severity of the disease. One reason is that, even under the traditional thresholds of statistical significance, drugs with severe side effects still manage to receive regulatory approval [97]–[100]. Therefore, regulatory agencies mandated to protect the public's health are understandably reluctant to adopt more risk-tolerant statistical criteria for

drug approvals.

However, there is an inexorable trade-off between the risk of false positives and that of false negatives, so being risk averse with respect to one criterion necessarily means being risk tolerant with respect to the other criterion. BDA seeks to balance the risks of both criteria simultaneously by explicitly minimizing the expected loss to patients due to both Type I and II errors, where the expected loss is the weighted sum of the measured impact of false positives and false negatives, weighted by their probabilities. This minimization process yields optimal false-positive and false-negative rates that reflect the different costs and benefits of each type of error, yielding an outcome that offers the "greatest good for the greatest number."

Montazerhodjat *et al.* [15] find that the optimal alphas determined by Bayesian decision analysis (BDA) were often much larger than 2.5% for terminal cancers with short survival times and no effective therapies, such as glioblastoma (a 47.5% BDA-optimal type 1 error), and smaller than 2.5% for less serious cancers with long survival times and multiple effective therapies (e.g., a 0.9% BDA-optimal type 1 error for early stage prostate cancer). Isakov *et al.* [14] provide corresponding results for the 25 most lethal diseases in the United States.

Chaudhuri *et al.* [27] apply Bayesian patient-centered models to anti-infective therapeutics, incorporating epidemiological models to determine the optimal alpha during outbreaks of epidemic disease. Most recently, a survey of over 2,700 Parkinson's disease (PD) patients by Hauber *et al.* [13] find that risk thresholds in a BDA framework for new neurostimulative devices in the treatment of PD increase markedly with the perceived benefit of the device to the patient. BDA has also been applied retrospectively to medical devices for treating obesity [72], to adaptive platform trials [73], and is being considered as a prospective input to the trial design of devices to treat kidney disease [17].

BDA applications do require more information than the traditional approach—the losses under both types of errors must be specified, and in some cases these losses may be difficult to gauge. However, several metrics have been developed for this purpose in the health technology assessment literature, including quality adjusted life

years (QALYs) to assess burden of disease, and sophisticated econometric survey tools designed by patient advocacy groups to measure the preferences of their constituents.

The most challenging practical issue in implementing this framework is the consequences of a larger number of false positives. This can be addressed by creating a temporary license to market "speculative" therapies that expires after a short period (say, two or three years), as described in [14]. During this period, the licensee is required to collect and share data on the performance of its therapy, and if the results are positive, the license converts to a standard approval, otherwise the therapy is withdrawn upon expiration. Regulators should have the right to terminate the temporary license at any time in response to adverse events or significantly negative data. Such licenses would greatly accelerate the pace of therapeutic development for many underserved medical needs without limiting regulatory flexibility.

Flexibility is particularly important because any system can be gamed, leading to unintended outcomes, hence no single interest group should be allowed to exercise undue influence in this process. Therefore, regulators must, and do, apply discretion, judgment, and a wealth of experience in their review process. Nevertheless, a systematic, rational, transparent, reproducible, and practical framework in which regulators' decisions can be clearly understood by and communicated to all stakeholders while explicitly incorporating their feedback may still have value.

5.3 Methods

We adopt the BDA framework proposed in [14] and calibrate its parameters using the available data for ALS and AMX0035. The assumed values of the parameters are listed in Table 5.2. To calibrate our BDA model to match the phase 2 trial design of AMX0035 [38], we assume an imbalanced two-arm randomized clinical trial in which participants are randomly assigned to treatment and control arms with a ratio of 2:1. We denote the size of the treatment and control arms by $2n$ and n, respectively. The Bayesian cost matrix has the same structure as in [14], and is shown in Table 5.1. Here, $N = 16,583$ denotes the estimated prevalence of ALS in the U.S. in 2015

Table 5.1: Cost matrix of Bayesian decision analysis.

	$\hat{H}=0$ (do not approve)	$\hat{H}=1$ (approve)	In-Trial Loss
$H=0$ (no effect)	0	Nc_1	$2nc_1$
$H=1$ (effective)	Nc_2	0	$n\gamma Nc_2$

[90]. The cost of Type I error is proportional to the loss per patient c_1 due to the adverse effects of ALS treatment. We assume that $c_1 = 0.07$, the value used in [14], which accounts for the adverse effects of all medical treatments. This is likely to be an overestimate of c_1 for ALS for two reasons. First, most adverse effects reported in the phase 2 trial of AMX0035 are gastrointestinal events [38], which are milder than a number of adverse medical effects used to estimate c_1 (e.g., amputation of a limb, traumatic brain injury, etc.). In addition, there is abundant clinical evidence to support the safety of the two drugs (sodium phenylbutyrate and TUDCA) in the AMX0035 combination therapy. Sodium phenylbutyrate was approved by the FDA in 1996 to treat urea cycle disorders, and TUDCA was tested in a small phase 2 trial (with 34 patients) in 2012 to treat ALS with no significant adverse effects reported [101].

Similarly, the cost of Type II error is proportional to the loss due to the disease burden of ALS suffered by each patient c_2. We use the heuristic proposed in [14] to estimate c_2

$$c_2 = \frac{D + YLD}{D + N} \tag{5.1}$$

Here, D denotes the number of deaths caused by the disease, YLD is the number of years lived with disability, and N is the disease prevalence, measured in age-standardized values. So far, we have not identified any study in the literature that estimates the disease severity specifically for ALS. Instead, we use the corresponding values for all motor neuron diseases [102] (which includes ALS, spinal muscular atrophy, hereditary spastic paraplegia, primary lateral sclerosis, progressive muscular atrophy, and pseudobulbar palsy) as a proxy. The age-standardized values of the parameters are $D = 0.46$, $YLD = 1.0$ and $N = 4.5$ (per 100,000 individuals). Using

Equation 5.1, we have $c_2 = 0.29$. For comparison, the authors of [14] estimate the disease severity for brain cancer ($c_2 = 0.30$) and for leukemia ($c_2 = 0.21$).

We also consider an alternative definition of disease severity using disability-adjusted life years (DALY) instead of YLD in Equation 5.1. The motivation behind using DALY to estimate disease severity is to account for the physical and mental affliction of ALS patients caused by the exacerbation of muscular atrophy and respiratory paralysis over the span of three to five years. We find that the DALY of ALS is 13.2 while that of medical adverse effects is 53.1 (per 100,000 individuals) [14], [102]. Consequently, the disease severity estimate using DALY is $\widetilde{c}_1 = 0.16$ for medical adverse effects and $\widetilde{c}_2 = 2.75$ for ALS. Since the ratio of Type II versus Type I errors c_2/c_1 is much higher for the DALY estimates (16.81) than YLD estimates (4.37), we expect that the corresponding Bayesian optimal Type I error rates will be higher for DALY estimates as well.

We calibrate the signal-to-noise ratio ρ of the treatment effect, with the results reported in the AMX0035 phase 2 trial [38]. The mean outcome of the treatment arm, measured by the ALSFRS-R score, is $\hat{\mu}_t = 29.06$ (with 87 subjects and standard error $s_t = 0.78$), while that of the control arm is $\hat{\mu}_c = 26.73$ (with 48 subjects and standard error $s_c = 0.97$). Assuming independently and identically distributed (IID) outcomes in each arm, we compute the pooled standard deviation $\hat{\sigma} = 1.24$ and a signal-to-noise ratio $\hat{\rho} = (\hat{\mu}_t - \hat{\mu}_c)/\hat{\sigma} = 0.33$. We use the estimate $\rho = 0.33$ as the baseline value, and vary ρ in the sensitivity analysis.

In addition, the imbalanced randomization ratio of 2:1 to treatment and control arms should reflect the prior belief of the clinical researchers that the efficacy of AMX0035 is superior to the placebo prior to the initiation of phase 2 trial. This is supported by both the ethical principle of clinical equipoise [103] as well as the results of a previous phase 2 trial that shows the efficacy of TUDCA in slowing ALS disease progression [101]. This prior belief is naturally incorporated into our Bayesian framework by assigning a Bayesian prior probability p_0 that the drug is ineffective or has adverse effects and setting $p_0 < 50\%$. In our simulations, we set $p_0 = 30\%$ as the baseline value and vary p_0 to gauge the sensitivity of the results to this parameter.

Finally, we place a constraint on the maximum power (or equivalently, the minimum Type II error β_{min}) of testing the alternative hypothesis. This reflects the practical considerations of the pharmaceutical industry in designing a clinical trial where the patient sample size is calculated with a target power of 80% or 90%. In our simulations, we set $\beta_{min} = 20\%$ to match the target power in the statistical analysis plan for the phase 2 trial of AMX0035 [38].

Using the calibrated values of simulation parameters (Table 5.2), the algorithm to compute the BDA-optimal trial sample size and Type I error rate follows directly from [14] and is described in detail in Section C.1.

Table 5.2: Assumed values of parameters in the Bayesian clinical trial model.

Parameter	Value(s)	Description	Sources
N	16,583	Prevalence of ALS in the U.S. in 2015	[90]
c_1	0.07 (0.16)	Severity of adverse effects of ALS treatment calibrated by YLD (DALY)	[14]
c_2	0.29 (2.75)	Disease severity of ALS calibrated by YLD (DALY)	[14], [102]
p_0	[0.1, 0.3, 0.5, 0.7, 0.9]	Probability that ALS treatment is ineffective or has adverse effects. Adjusted to model different scenarios with baseline value $p_0 = 0.3$	[38]
ρ	[0.25, 0.33, 0.5]	Signal-to-noise ratio of treatment effect. Adjusted to model different scenarios with baseline value $\rho = 0.33$	[38]
γ	$4 \times 10^{-3}\rho$	Incremental cost incurred due to adding an extra patient to each arm	[14]
β_{min}	0.2	Minimum Type II error of BDA	[14], [38]

5.4 Results

Table 5.3 summarizes the results where the burden of disease measures c_1 and c_2 are calibrated using YLD in the top panel and corresponding results using DALY are in the bottom panel. To simulate different scenarios of treatment effects and prior probabilities of an effective drug, we simultaneously vary the values of ρ (the signal-to-noise ratio of treatment effects) and p_0 (the prior probability that the drug is ineffective).

For the YLD, the BDA-optimal Type I error rate α^* exceeds the p-value of 3% reported in the phase 2 trial of AMX0035 for all combinations of parameter values with $p_0 \leq 70\%$, except when $p_0 = 70\%$ and $\rho = 0.5$ (and the corresponding $\alpha^* = 2.9\%$). When the prior probability of an effective drug is at least 30%, the BDA-optimal type I error threshold suggests that approving AMX0035 maximizes the expected benefits to patients. When the prior probability of efficacy is 90%, the BDA-optimal threshold is 51.0%.

As expected, we find that α^* computed using the YLD estimates of disease severity is more conservative (i.e., lower) than the corresponding values computed using a DALY burden-of-disease measure, since the latter reflect the afflictions caused by the progression of ALS and have a higher cost ratio c_2/c_1. In fact, the bottom panel of Table 5.3 shows that all combinations of parameter values with $p_0 \leq 70\%$ yield BDA-optimal α^* higher than 10.7%, reflecting the urgency of the unmet medical needs of ALS patients.

Table 5.3 shows that the optimal sample size n^* decreases with higher values of ρ, which agrees with the findings of [14]. The optimal Type I error rate α^* also decreases with higher values of p_0, which is sensible since the expected cost due to Type I error increases with p_0 and the Bayesian analysis results in a more stringent (i.e., lower) α^* to lower the expected cost.

Table 5.3: Optimal sample size and Type I error rate for a hypothetical ALS therapy with randomization ratio 2:1.

p_0(%)	ρ	n^*	λ^*	α^*(%)	Power (%)
Burden of Disease Measure: YLD					
10	0.25	51	0.00	50.0	80.0
10	0.33	30	0.02	49.0	80.0
10	0.50	12	−0.03	51.0	80.0
30	0.25	249	1.02	15.4	80.0
30	0.33	141	1.02	15.3	80.0
30	0.50	63	1.03	15.2	80.0
50	0.25	387	1.48	7.0	80.0
50	0.33	222	1.50	6.7	80.0
50	0.50	99	1.50	6.6	80.0
70	0.25	513	1.83	3.4	80.0
70	0.33	297	1.87	3.1	80.0
70	0.50	135	1.90	2.9	80.0
90	0.25	645	2.21	1.3	79.0
90	0.33	399	2.30	1.1	80.0
90	0.50	186	2.37	0.9	80.0
Burden of Disease Measure: DALY					
10	0.25	3	−0.64	73.8	80.0
10	0.33	3	−0.57	71.5	80.0
10	0.50	3	−0.43	66.8	80.0
30	0.25	50	0.00	50.0	80.0
30	0.33	30	0.02	49.2	80.0
30	0.50	12	−0.03	51.0	80.0
50	0.25	168	0.69	24.6	80.0
50	0.33	96	0.70	24.3	80.0
50	0.50	42	0.69	24.6	80.0
70	0.25	306	1.22	11.1	80.0
70	0.33	174	1.23	10.9	80.0
70	0.50	78	1.24	10.7	80.0
90	0.25	513	1.83	3.4	80.0
90	0.33	297	1.87	3.1	80.0
90	0.50	135	1.90	2.9	80.0

To provide a more direct analysis of the AMX0035 trial, we compute the BDA-optimal Type I error rate α^* while setting the sizes of the treatment arm ($n_1 = 89$) and control arm ($n_2 = 48$) to match the actual sizes of the phase 2 trial of AMX0035. Table 5.4 summarizes the results where the disease severities c_1 and c_2 are calibrated using YLD and DALY under different assumptions for p_0 and ρ.

For both YLD and DALY measures, the BDA-optimal α^* are higher than 10% for all scenarios where $p_0 \leq 70\%$ and $\rho < 0.5$. The differences between the two burden-of-disease measures are smaller in this case because we have fixed the trial size, hence the cost ratio c_2/c_1 is less significant. In particular, in the baseline model, where we assume $p_0 = 30\%$ and $\rho = 0.33$, the optimal $\alpha^* = 15.4\%$ for both estimation methods, four times higher than the reported p-value of 3%. When the signal-to-noise ratio is $\rho = 0.5$, we find that $\alpha^* = 2.6\%$, (close to the reported p-value) for all p_0 between 10% and 90%.

In summary, our BDA analysis supports approving AMX0035 in all scenarios except when $p_0 = 90\%$, which is implausible given the abundant clinical evidence on the safety and efficacy of the two drugs in AMX0035 [38], [101], and is also inconsistent with the presumption of clinical equipoise for the clinical trial.

Table 5.4: Optimal Type I error rate for phase 2 trial of AMX0035 with 89 patients in the treatment arm and 48 in the control arm.

p_0(%)	ρ	λ^*	α^*(%)	Power (%)
Burden of Disease Measure: YLD				
10	0.25	0.55	29.0	80.0
10	0.33	1.02	15.4	80.0
10	0.50	1.95	2.6	80.0
30	0.25	0.55	29.0	80.0
30	0.33	1.02	15.4	80.0
30	0.50	1.95	2.6	80.0
50	0.25	0.63	26.3	77.7
50	0.33	1.02	15.4	80.0
50	0.50	1.95	2.6	80.0
70	0.25	1.24	10.7	56.1
70	0.33	1.18	11.8	75.1
70	0.50	1.95	2.6	80.0
90	0.25	2.21	1.4	20.8
90	0.33	1.91	2.8	48.1
90	0.50	1.95	2.6	80.0
Burden of Disease Measure: DALY				
10	0.25	0.55	29.0	80.0
10	0.33	1.02	15.4	80.0
10	0.50	1.95	2.6	80.0
30	0.25	0.55	29.0	80.0
30	0.33	1.02	15.4	80.0
30	0.50	1.95	2.6	80.0
50	0.25	0.55	29.0	80.0
50	0.33	1.02	15.4	80.0
50	0.50	1.95	2.6	80.0
70	0.25	0.55	29.0	80.0
70	0.33	1.02	15.4	80.0
70	0.50	1.95	2.6	80.0
90	0.25	1.24	10.7	56.1
90	0.33	1.19	11.8	75.0
90	0.50	1.95	2.6	80.0

5.5 Discussion

The BDA framework formalizes the notion that potentially effective treatments for terminal diseases with no effective treatments such as ALS should be evaluated with a higher p-value threshold than the traditional 2.5% or 5% value. This conclusion also aligns with the patient preferences expressed by ALS advocacy groups [94]. The optimal Type I error of the baseline model of $\alpha^* = 15.3\%$, using the conservative YLD estimates for ALS severity, is between the corresponding values for lung cancer (13.7%) and pancreatic cancer (23.9%) found in [14], a reasonable reflection of the prevalence and severity of ALS.

Two limitations of our analysis need to be addressed in future work. First, a more accurate and rigorous procedure is needed to estimate the loss of Type II error (i.e., ALS disease severity c_2). The results of our Bayesian decision analysis are highly sensitive to disease severity estimates, hence input from ALS medical experts and patient advocates should be incorporated into the final values used by decision makers, as proposed by [13]. Our heuristic to estimate c_2 (Equation 5.1) uses the mortality rate and years lived with disability (YLD) of ALS, while the phase 2 trial of AMX0035 reports the reduction in the ALS disease progression (measured by the motor function scale named ALSFRS-R) as its primary outcome [38]. This reduction in ALS progression must be translated into an equivalent reduction in YLD or DALY to accurately gauge the patient benefits of an effective treatment.

In addition, we assume that the treatment outcomes for the patients are IID in each arm of the clinical trial. In the actual phase 2 trial of AMX0035, the average treatment effect is estimated by fitting the treatment outcome with a mixed-effects model, which accounts for the variation in treatment outcomes due to covariates such as age and the previous stage and rate of ALS progression for each patient [38]. Without the knowledge of the covariates and the treatment outcome of each patient, it is impossible to reproduce the procedure of the actual phase 2 trial to replicate the reported p-value of 3%. As a result, we assume IID treatment outcomes and vary the signal-to-noise ratio ρ in the sensitivity analysis to account for potential

miscalibrations of our model in estimating the treatment effect.

5.6 Conclusion

Based on the phase 2 trial results of AMX0035, the benefits of therapeutic effects outweigh the relatively minor risks of adverse effects from the BDA perspective. That perspective strikes an optimal tradeoff between Type I and Type II errors, yielding a BDA-optimal p-value threshold which, under a wide range of realistic assumptions, is consistently higher than the reported value of 3% of the trial data. While we recognize the complexity of factors involved in the regulatory decision, from the perspective of maximizing the benefits to the ALS patient community the therapeutic effects observed in the phase 2 study strongly support the regulatory approval of AMX0035.

Part IV

Data Analytics for Drug Development Forecasting

Chapter 6

Identifying and Mitigating Potential Biases in Predicting Drug Approvals

Machine learning models have increasingly been applied to predict the drug development outcomes of novel therapeutics based on intermediary clinical trial results, reducing the significant financial risks of drug development. A key challenge to the correct prediction of the drug development outcome is the presence of various forms of bias in the historical drug approval data and the prediction model. By instantiating the Debiasing Variational Autoencoder (DB-VAE), the state-of-the-art deep learning model for automated debiasing, we simultaneously identify the bias in both the drug approval outcomes and drug features and mitigate the bias in the model's predictions. We find that the debiased model improves the prediction performance with higher true positive rates and F_1 scores than its un-debiased counterparts. We also show that the debiased model generates additional financial value for the drug developer in six major therapeutic areas, ranging from $763 to $ 1365 million. Our analysis illustrates the importance of debiasing in improving financial efficiency and reducing the financial risks of late-stage drug development. [1]

[1]Joint work with Elaheh Ahmadi, Alexander Amini, Daniela Rus, and Andrew W. Lo. This chapter was published as a research article in *Drug Safety* [28]

6.1 Introduction

6.1.1 Financial risks in novel drug development

Despite groundbreaking advances in biomedical science and technology, the translational research and development (R&D) of novel therapeutics has become more expensive and less likely to succeed over the last five decades. The number of new drugs approved by the U.S. Food and Drug Administration (FDA) per billion U.S. dollars ($) spent on translational R&D has halved in inflation-adjusted terms about every 9 years since 1950 [1]. A recent analysis [4] estimates a median cost of $985.3 million to bring a new drug to market in the period 2009 to 2018, while the historical probability of success (PoS) of developing a novel drug from a phase 1 clinical trial to FDA approval is merely 10.8% in all therapeutic areas, and as low as 4.0% in oncology [3]. Novel scientific and business models are needed to reduce the financial risks and bridge the funding gap (also known as the "valley of death" [7]) in the translational R&D of novel therapeutics.

6.1.2 Machine learning for drug approval prediction

In recent years, machine learning and artificial intelligence (AI) have been increasingly applied to forecast the PoS of FDA approval for a novel drug candidate based on early or mid-stage (phase 1 or 2) clinical trial results. Early works [18]–[20], [104] revealed important factors which correlated with successes and failures. However, these conclusions have been limited by the relatively small size of the datasets, which consist of fewer than 100 drugs or 500 clinical trials. Beinse *et al.* [40] trained a regularized Cox model with 462 anti-neoplastic agents to predict drug approval from phase 1 results, although their model is affected by look-ahead bias, since the authors randomly split the data into training and testing sets, training their model with future data but evaluating it with past data. Lo *et al.* [21] were the first to train machine learning models on the Citeline Informa dataset [22], with more than 91,000 drugs and 374,000 clinical trials . Using the Informa dataset, Wong *et al.* [2] proposed a

path-by-path approach to estimate the PoS of drug approvals in different therapeutic areas, which is robust against missing data. Recently, Siah *et al.* [24] organized a data science competition with over 50 participating teams to predict drug approvals using the Informa dataset. The top-performing models in this competition used novel features that are highly predictive of drug approval outcomes.

A key challenge to predicting the approval outcomes of drug development projects is to address the various forms of bias present in the dataset and prediction model. Two major sources of bias are data missingness and dataset imbalance. While data missingness can be addressed by imputation [21], dataset imbalance (i.e., over- /under-representation in outcome labels and input features) remains a critical challenge that limits the predictive performance of machine learning models (Figure 6-1). In its outcome labels, only 11.8% of all drugs in our dataset are approved by the FDA. Thus, the prediction model trained on the imbalanced dataset is incentivized to predict negative outcomes, and has a low true positive rate. In the input feature space, imbalance occurs where many drugs have similar properties (e.g., "me too" [105] or repurposed drugs with biochemical features similar to previously approved ones). Since drugs with similar properties are not guaranteed to be effective to treat different diseases, the prediction model has a lower performance when predicting the approval outcomes of overrepresented drugs in the feature space.

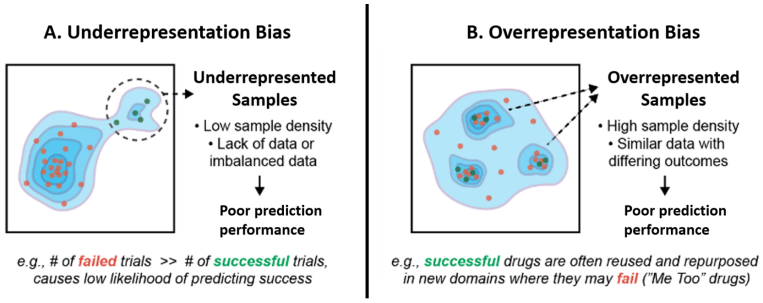


Figure 6-1: Sources of bias in Informa dataset of drug development. (A) Underrepresentation in outcome labels with 11.8% of positive samples (green); (B) Overrepresentation in drug and clinical trial features (e.g., "me too" or repurposed drugs with similar drug features as previously approved drugs).

6.1.3 Mitigating the bias of machine learning models

With the increasing reliance on machine learning models to automate the decision-making process in many applications, the issue of bias and fairness of machine learning models has become a central concern (see [41] for a systematic survey). Previous studies reported and addressed significant racial, gender and socioeconomic biases in machine learning models applied to domains such as financial loans [106], criminal justice [107], career advertisement [108], medical diagnosis [109], [110], and healthcare policy [111].

Bias in a machine learning model is often caused by biases within an imbalanced training dataset [41]. The biased model performs better on overrepresented subgroups in the dataset than on underrepresented minorities [42], [112], [113]. Common strategies used to mitigate algorithmic bias include rebalancing the training dataset by resampling the training data [114], [115], generating synthetic data [116], [117], or changing a subset of the class labels [107]. Each method, however, has its limitations. Resampling requires knowledge of the class imbalance, and is difficult when the class label is latent (e.g., gender or skin color in an image) and needs to be manually annotated. Zhou and Liu [114] showed that resampling methods that are effective for binary classification often do not generalize to multi-class classification. Generating synthetic data requires modeling the distribution of input features, which may be difficult for high-dimensional features, and may produce unrealistic synthetic samples. Recently, Bandi and Bertsimas [107] proposed an optimization framework which flips a subset of binary labels to achieve guaranteed demographic parity, although the model does not justify why certain labels are flipped, and it is evaluated on a testing set which contains samples with flipped labels.

Given these challenges, Amini *et al.* [42] proposed a novel framework called the Debiasing Variational Autoencoder (DB-VAE), which automatically identifies and mitigates the bias in large datasets without prior knowledge of the detailed structure of bias (e.g., whether certain subgroups are over/underrepresented). The mathematical framework and debiasing mechanism of DB-VAE are discussed in Section 6.3.2.

6.1.4 Contributions of this work

In this work, we simultaneously identify and mitigate various forms of bias by instantiating the Debiasing Variational Autoencoder (DB-VAE) for drug approval predictions. The automatic debiasing feature of DB-VAE is particularly useful in our application, since it is difficult to directly analyze the bias structure of complex drug and clinical trial data. The main contributions of our work are:

1. To the best of our knowledge, this work is the first to systematically address the significant bias present in the machine learning model for drug approval prediction using one of the largest datasets in this domain.

2. We show that the debiased model achieves better overall prediction performance and its prediction performance is more uniform across different subgroups of the dataset than its un-debiased counterpart.

3. We quantify the impact of different forms of bias on the prediction performance. Debiasing the imbalance of drug approval outcomes results in major improvements for all drugs, while debiasing the input feature distributions achieves improvements for drugs which are overrepresented in the input feature space, such as oncology and cardiovascular drugs.

4. We show that debiasing generates significant financial values of drug development in six major therapeutic areas.

5. From the pharmacovigilance perspective, we find that the debiased model predicts safe and effective drugs more accurately than its un-debiased counterpart.

6.2 Data

We query historical drug development data in the period from 2004 to 2020 (inclusive) in the Citeline Informa database [22]. Since a drug developer may conduct multiple clinical trials for one drug to investigate its therapeutic efficacy for different diseases,

we predict the binary outcome of whether the drug has been approved by the FDA to treat a particular indication, which we call a "drug-indication pair." For clarify of exposition, we shall refer to "drug-indication pair" simply as "drug" in the subsequent sections. We follow the data query and preprocessing procedures described in [21], and refer the reader to this work for additional details. We form the Phase 2 to Approval (P2APP) dataset, which consists of drug-indication pairs with known approval outcomes (success or failure) and their phase 2 clinical trial results. We train the machine learning model on P2APP to predict the binary approval outcome from drug features and phase 2 trial results. The summary statistics of the P2APP dataset are provided in Table 6.1. The raw data consists of both categorical and continuous features. A categorical feature may be single-labeled (e.g., whether the drug was previously approved for another indication) or multi-labeled (e.g., the drug developer may conduct clinical trials in different countries for a drug). We apply one-hot encoding to the multi-labeled features and create binary child features. Detailed descriptions of the drug and clinical trial features are summarized in Table D.1. In addition, due to the different standards of post-study reporting of clinical trial results (especially before the 2007 FDA Amendments Act) [21], there is considerable missingness in certain drug and clinical trial features (Table D.2) which needs to be imputed. We discuss the details of imputation procedure in Section 6.3.3.

Table 6.1: Summary statistics of P2APP dataset.

Dataset[2]	Type	Drugs	Indications	Clinical Trials	Drug-Indication Pairs
P2APP (2004-2020)	Approved	685	190	2320	876
	Failed	3615	292	10288	6555
	Total	4079	298	12397	7431
Training Set (2004-2018)	Approved	595	182	2060	752
	Failed	3200	275	8090	5630
	Total	3612	283	10035	6382
Testing Set (2019-2020)	Approved	108	81	344	124
	Failed	637	195	2397	925
	Total	740	212	2711	1049

The training set consists of drug approval outcomes from 2004 to 2018 (both inclusive) and testing set from 2019 to 2020. A drug may be approved to treat an indication but fail in other indications. Therefore, the sum of numbers of approved and failed drugs is not equal to the total number of unique drugs. However, the sum of numbers of approved and failed drug-indication pairs is equal to the total number of unique drug-indication pairs.

6.3 Methods

6.3.1 Algorithm fairness

We adopt the definition of algorithm fairness proposed in [42]. Given the training dataset $D_{train} = \{(x^{(i)}, y^{(i)}) : 1 \leq i \leq n\}$, where $x^{(i)} \in X \subset \mathbb{R}^m$ denotes the m-dimensional feature vector and $y^{(i)} \in Y = \{0, 1\}$ denotes the binary prediction target, our goal is to find a classifier $f : X \to Y$ which is fair with respect to sensitive features $z \in \mathbb{R}^d$. The sensitive features z may either be observed in D_{train} (e.g., the outcome of drug development) or latent, which means that $z = z(x)$ are not directly observed, but can be represented as a function of the observed features x. For example, in the computer vision task of face recognition, $x^{(i)}$ is the input image, $y^{(i)}$ denotes whether the image contains a human face, and the latent features $z^{(i)} = z(x^{(i)})$ include skin color, gender, and the age of the subject in the image. We

define an unbiased classifier f with respect to the sensitive features z if $f(x) = f(x, z)$, i.e., the classification decision is not affected by the additional sensitive features, and a biased classifier if $f(x) \neq f(x, z)$ for some z.

As pointed out in [42], in order to train an unbiased classifier, the training samples in D_{train} should be uniformly distributed across the latent feature space Z. Furthermore, given a classifier f and testing dataset D_{test}, we can measure the bias of f by the variance of its prediction performance across different subgroups in Z. A larger variance indicates greater bias, since the classifier performs more poorly on certain subsets than others. For our purposes, the sensitive features of interest z are the drug approval outcomes (observed) and the degree of over/underrepresentation of a drug in the feature space (latent).

6.3.2 Debiasing via DB-VAE

We instantiate the Debiasing Variational Autoencoder (DB-VAE) [42] to train a debiased classifier for drug approval prediction. DB-VAE consists of a Variational Autoencoder (VAE) [118], which learns the latent features $z(x)$ and predicts the approval outcome $\hat{y} = f(x)$, coupled with a feedback loop which debiases the training dataset by adaptively adjusting the resampling weights for over- and underrepresented samples. The model architecture of DB-VAE is shown in Figure 6-2. For clarify of presentation, we review the key components of DB-VAE that are most relevant to our application, and refer the reader to the original work [42] for additional details.

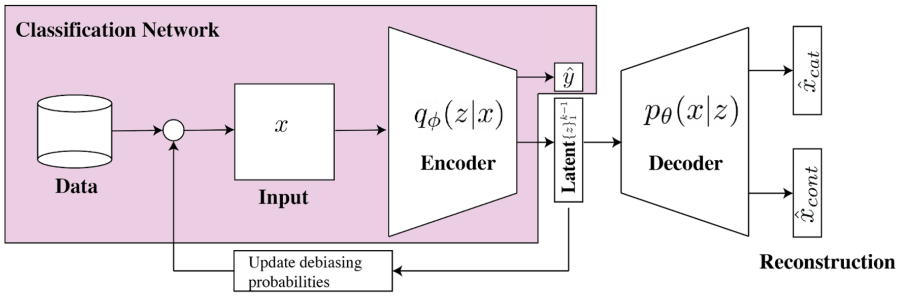


Figure 6-2: Architecture of debiasing variational autoencoder (DB-VAE) instantiated for predicting drug development outcomes.

We train DB-VAE to simultaneously debias the imbalanced drug approval outcomes and the over/underrepresentation in the input feature space. To debias the outcome labels, we enforce that each training batch (of size 32) of the stochastic gradient descent contains an equal number of positive and negative training samples, which ensures that the model is not biased to predict negative outcomes due to overrepresentation in outcome labels.

To debias the input features, we utilize the low-dimensional latent representation $z(x) = [z_1(x), ..., z_d(x)]$ of the high-dimensional features $x \in \mathbb{R}^m$ (with $d \ll m$) learned by the VAE with an encoder-decoder architecture. The encoder network produces the latent features $z(x)$ as its output. The decoder network then reconstructs the input features $\hat{x} = \hat{x}(z(x))$ from $z(x)$. To ensure that the reconstructed features are close to their original values $\hat{x} \approx x$, we use L_1 loss function for reconstructing continuous features (denoted as $L_{r,\ con}$) and a weighted binary cross entropy loss (denoted as $L_{r,\ cat}$ in Equation 6.1) for reconstructing categorical features.

$$L_{r,\ \ cat} = -\frac{1}{N} \sum_{c \in cat} \sum_{i=1}^{N} \frac{1 - \bar{x}_c}{\bar{x}_c} x_c^{(i)} \log \hat{x}_c^{(i)} + (1 - x_c^{(i)}) \log (1 - \hat{x}_c^{(i)}) \tag{6.1}$$

Here $x_c^{(i)}$ denotes the value of categorical feature c of the i-th training sample, $\hat{x}_c^{(i)}$ denotes the reconstructed value by DB-VAE, and $\bar{x}_c$ denotes the average value of feature c in the training dataset. This weighted loss achieves a more accurate reconstruction for categorical features which are exceedingly sparse or dense. For sparse features (with $\bar{x}_c$ close to 0), we assign a higher weight to the positive samples ($x_c^{(i)} = 1$) and increase the true positive rate of reconstruction. For dense features (with $\bar{x}_c$ close to 1), we assign a higher weight to the negative samples ($x_c^{(i)} = 0$) and increase the true negative rate of reconstruction.

In addition, VAE uses the Kullback–Leibler (KL) divergence [119] to regularize the latent space distribution and prevent overfitting. We denote this regularization loss by L_{KL}. As in [42], we use the encoder to predict the PoS of the drug approval outcome $\hat{y} = z_0(x) \in (0, 1)$, and denote the cross-entropy loss by L_c. The predicted PoS $\hat{y}$ is not used in debiasing.

Given the latent representation $z(x)$ of every sample x in the training set X_{train}, we compute the probability density function (PDF) $Q_i(z_i(x)|X_{train})$ of each latent dimension $i \in 1, ..., d$ via a histogram with 10 bins. We use the notation of conditional probability $Q_i(\cdot|X_{train})$ to emphasize that the latent space density is sensitive to the distribution of input features in the training dataset X_{train}. The joint PDF in the d-dimensional latent space is $Q(z(x)|X_{train}) = \prod_{i=1}^{d} Q_i(z_i(x)|X_{train})$. Based on the latent space density $Q(z(x)|X_{train})$, [42] proposed a debiasing algorithm which assigns higher probabilities of resampling to training samples with lower $Q(z(x)|X_{train})$ (i.e., that are underrepresented in the latent space) into the next training batch and assigns lower resampling probabilities to the overrepresented samples. Specifically, the debiasing resampling weight $W(z(x)|X_{train})$ for training sample x is given by

$$W(z(x)|X_{train}) \propto \prod_{i=1}^{d} \frac{1}{Q_i(z_i(x)|X_{train}) + \alpha} \tag{6.2}$$

The proportionality sign $\propto$ indicates that the sum of $W(z(x)|X_{train})$ is normalized to 1. The debiasing smoothing parameter $\alpha > 0$ controls the degree of debiasing. A smaller value of α corresponds to more aggressive debiasing, since $W(z(x)|X_{train})$ is mostly determined by the latent space density $Q_i(z_i(x)|X_{train})$. On the other hand, a large value of $\alpha \gg \max_{i,x} Q_i(z_i(x)|X_{train})$ corresponds to uniform resampling, and does not debias the latent space distribution. The model parameters are trained by minimizing the loss function

$$L_{DB-VAE} = \lambda_1 L_c + \lambda_2 L_{r,\ cat} + \lambda_3 L_{r,\ con} + \lambda_4 L_{KL} \tag{6.3}$$

where the weights $\lambda_i > 0$ are model hyperparameters. We choose the default values $\lambda_1 = 10, \lambda_2 = \lambda_3 = 1$ and $\lambda_4 = 0.001$ to reflect the relative importance of each term in the loss function L_{DB-VAE}. Sensitivity analysis (Table D.4) shows that the model performance is robust against a wide range of hyperparameter values.

6.3.3 Training DB-VAE

Since drug development is a non-stationary process in which drugs approved in the past set higher standards for drug candidates in the future, there will be a significant look-ahead bias if we randomly split the dataset into training and testing sets, training the models on future data but evaluating on past data. To avoid this look-ahead bias, we train the models with historical data from 2004 to 2018, and evaluate the models on out-of-sample data from 2019 to 2020. The model parameters are optimized by minimizing the loss function (Equation 6.3) via stochastic gradient descent (SGD) with 200 training epochs, batch size 32, and learning rate 10^{-5}. To determine the training epochs and learning rate, we observe the evolution of the classification and reconstruction losses ($L_c, L_{r,\ cat}, L_{r,\ con}$) on a held-out validation set, formed by sampling 10% of the training set randomly without replacement, and terminate SGD when the losses converge. We impute the missing entries using 5-nearest neighbor imputation [120] for training, testing, and validation sets separately. We apply a log-transform to the continuous features, and normalize each continuous feature to zero mean and unit variance. Since most input features are sparse, we only use those features whose variance is above 0.2 before imputation. We implement the DB-VAE model in TensorFlow. The model configuration and hyperparameter values are summarized in Table D.3. We perform sensitivity analysis on the results against different values of model hyperparameters. The results are summarized in Table D.4.

To quantify the contributions from the two forms of bias to the prediction performance, we train four instantiations of DB-VAE, which differ by whether the model debiases the imbalance of outcome labels in each training batch (DB-Label) and whether it debiases the distribution of latent representations of input features (DB-Latent). The nomenclature is summarized in Table 6.2, and used in the subsequent sections.

Table 6.2: Nomenclature of DB-VAE models.

Debiasing mechanism	Original outcome labels	Debiased outcome labels
Original latent space distribution	No-DB-Label, No-DB-Latent	DB-Label, No-DB-Latent
Debiased latent space distribution	No-DB-Label, DB-Latent	DB-Label, DB-Latent

6.4 Results

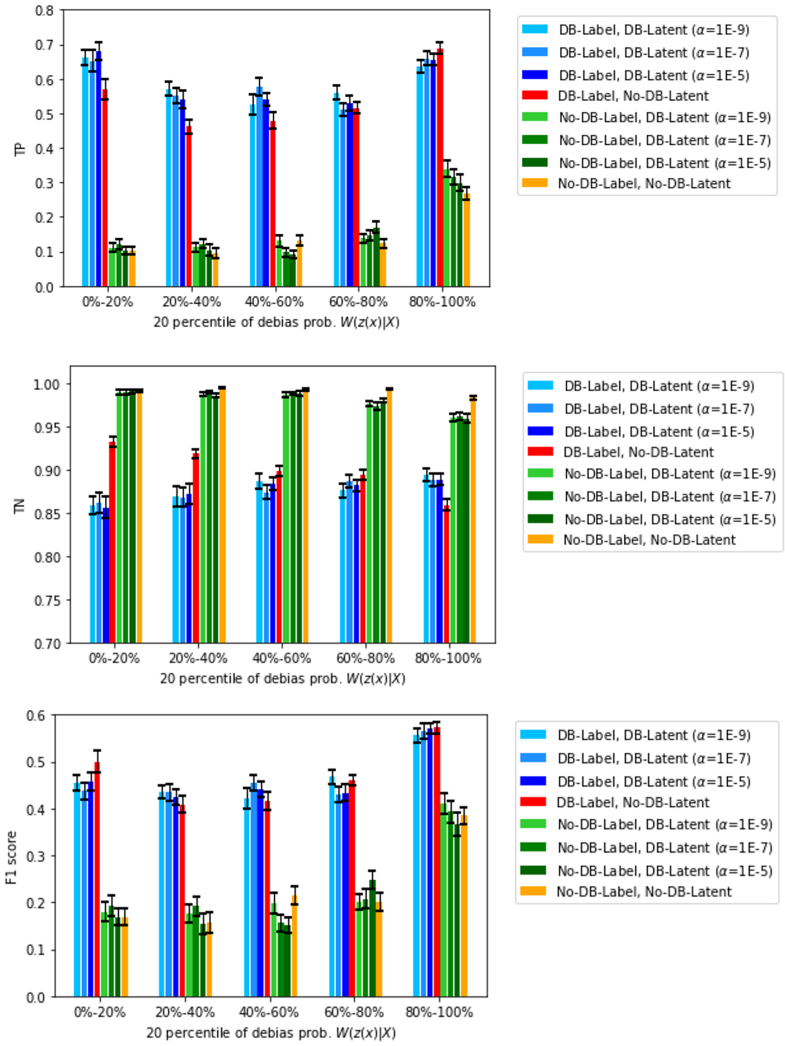
6.4.1 Prediction performance

We evaluate the prediction performance of the trained models on drug development outcomes from 2019 to 2020. We compute the confusion matrix for binary classification, and report the true positive (TP) rate, true negative (TN) rate, and F_1 score of each classifier. To quantify the uncertainty from imputing the missing entries, we train 30 instances of a given set of model hyperparameters, each with randomly split training and validation sets, and report the average value of their performance metrics and associated standard errors. We also report the area under receiver operating characteristic curve (AUC) of each model, which needs to be interpreted with care due to the imbalance in drug approval outcomes (Figure D-1) [121].

The prediction performance is summarized in Table 6.3. Comparing each DB-Label model with its No-DB-Label counterpart, we find that debiasing the outcome labels significantly improves both the TP and F_1 score in all therapeutic areas. Debiasing the latent space distribution (DB-Latent) improves the TP in three therapeutic areas (oncology, cardiovascular, and central nervous system), with a slightly lower TN. The tradeoff between higher TP and lower TN is consistent with the rationale of DB-Latent to achieve more uniform accuracy between different subgroups (approved vs. failed drugs) of the dataset [42]. The performance of DB-Latent and No-DB-

Latent are similar for autoimmune/inflammation, metabolic, and infectious disease. We conclude that the bias of imbalanced drug approval outcomes is a more severe issue which limits the prediction performance than the bias of over/underrepresentation in the input feature space.

To analyze the effect of debiasing on over/underrepresented drugs in the feature space, we evaluate the DB-VAE models on drugs in each quintile of $W(z(x)|X_{test})$. The results are shown in Figure 6-3. For each DB-Latent model, we use three different values of smoothing parameter α and the results are robust against different values of α. We observe that DB-Latent improves TP over its No-DB-Latent counterpart by 9.4%, 9.3%, 6.8% and 1.8% in the lowest four quintiles, with the greatest improvement in the two lowest quintiles (i.e., top 40% most overrepresented drugs in the test set). Debiasing the latent space distribution helps predict successful drugs which are overrepresented in the feature space (e.g., "me too" or repurposed drugs). This has major implications on the financial value of drug approval prediction, as will be shown in Section 6.4.4. The standard deviation of TP across the five quintiles is consistently lower for the DB-Label, DB-Latent models (5.1%, 5.8%, 6.5%) than their No-DB-Latent counterpart (8.2%), which confirms that DB-Latent reduces algorithmic bias across over/underrepresented subgroups.



Performance metrics (a) true positive rate, (b) true negative rate, and (c) F_1 score. For each DB-Latent model, we use three different values of smoothing parameter α and the results are robust against different values of α.

Figure 6-3: Effects of debiasing the drug approval outcome labels (DB-Label) and debiasing latent space distributions (DB-Latent) on prediction performance.

Table 6.3: Prediction performance of different instantiations of DB-VAE.

Therapeutic Area[3]	DB-Label	DB-Latent	F_1 score	SE	TP	SE	TN	SE
All	Yes	Yes	0.48	0.005	0.60	0.012	0.88	0.006
	Yes	No	0.49	0.004	0.57	0.005	0.90	0.002
	No	Yes	0.25	0.012	0.17	0.011	0.98	0.002
	No	No	0.25	0.007	0.15	0.005	0.99	<0.001
Oncology	Yes	Yes	0.35	0.007	0.57	0.017	0.88	0.007
	Yes	No	0.36	0.006	0.45	0.009	0.93	0.002
	No	Yes	0.23	0.015	0.18	0.015	0.98	0.003
	No	No	0.22	0.012	0.13	0.008	1.00	<0.001
Cardiovascular	Yes	Yes	0.42	0.031	0.52	0.041	0.87	0.010
	Yes	No	0.39	0.014	0.46	0.019	0.87	0.005
	No	Yes	0.17	0.047	0.12	0.033	0.99	0.003
	No	No	0.08	0.035	0.05	0.025	1.00	<0.001
Central Nervous System	Yes	Yes	0.60	0.007	0.63	0.013	0.90	0.005
	Yes	No	0.60	0.006	0.61	0.009	0.91	0.003
	No	Yes	0.20	0.021	0.12	0.015	0.99	0.002
	No	No	0.17	0.012	0.10	0.008	0.99	0.002
Autoimmune/ Inflammation	Yes	Yes	0.49	0.005	0.53	0.010	0.87	0.006
	Yes	No	0.49	0.005	0.52	0.007	0.88	0.003
	No	Yes	0.29	0.015	0.18	0.012	0.98	0.002
	No	No	0.33	0.011	0.21	0.008	0.99	0.001
Metabolic	Yes	Yes	0.56	0.011	0.66	0.019	0.81	0.012
	Yes	No	0.57	0.008	0.68	0.011	0.82	0.008
	No	Yes	0.26	0.017	0.17	0.015	0.97	0.005
	No	No	0.29	0.012	0.19	0.009	0.97	0.002
Infectious Disease	Yes	Yes	0.53	0.009	0.65	0.011	0.82	0.012
	Yes	No	0.52	0.008	0.67	0.010	0.80	0.010
	No	Yes	0.31	0.018	0.21	0.015	0.98	0.002
	No	No	0.29	0.016	0.19	0.012	0.98	0.001

Performance measured by F_1 score, true positive (TP) rate, and true negative (TN) rate.
Both models with DB-Latent use smoothing parameter $\alpha = 10^{-7}$.

6.4.2 Feature importance

We use the saliency score [122] to identify the input features which are the most important for predicting drug approval. The saliency score of a feature is a real number whose magnitude reflects the sensitivity of the model predictions to changes in the feature value and is commonly used to measure feature importance in deep learning models. Table 6.4 lists the top ten features of the DB-Latent, DB-Label model which have the highest absolute saliency scores. Some of these features (prior approval of the drug for another indication, whether the phase 2 trial meets the positive/negative endpoints, whether the delivery medium is powder) were identified as important features in previous work [21]. Our debiased model also reveals previously unidentified therapeutic factors such as the two pharmacology families (inducing cancer cell apoptosis and insulin-like growth factor receptor antagonist) and one biological target (ion channel). The model prediction is also sensitive to the year when the phase 2 trial is completed and the year when drug approval outcome is known, which reflects the non-stationarity of the drug development process.

Table 6.4: Top 10 drug and clinical trial features of DB-Label, DB-Latent model with the highest magnitudes of saliency scores (measured in 10^{-5}).

Feature	Saliency
Year of phase 2 trial completion	−2.16
Trial outcome - Completed, positive outcome/primary endpoint(s) met	1.72
Pharmacology - Induce cancer cell apoptosis	−1.69
Year of drug approval outcome	1.43
Biological Target - Ion channel	−1.42
Trial outcome - Terminated, lack of efficacy	−1.33
Medium - Powder	1.15
Pharmacology - Insulin-like growth factor receptor antagonist	1.09
Prior approval of drug for another indication	−1.00
Trial outcome - completed, negative outcome/primary endpoint(s) not met	−0.91

6.4.3 Latent space clusters

The encoder of DB-VAE learns a low-dimensional representation $z(x)$ which captures the structure of the high-dimensional distribution of input features x. The density of latent space distribution is then used in debiasing. To interpret the latent space of DB-VAE, we visualize the 10-dimensional latent representations of drugs in the testing set in 2 dimensions using t-SNE [123], and observe that the latent space of DB-VAE consists of two distinct clusters (Figure 6-4).

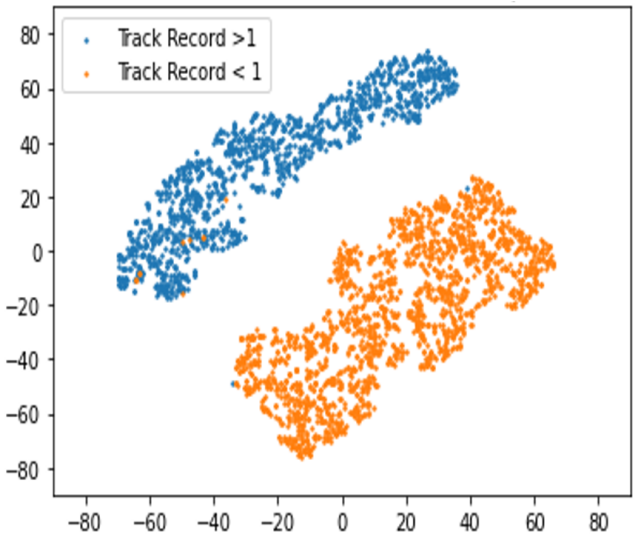


Figure 6-4: t-SNE visualization of the latent representation of DB-Label, DB-Latent model with smoothing parameter $\alpha = 10^{-7}$. Drugs in the two clusters are well separated by the values of track record for the clinical trial sponsors.

The drugs of the two clusters of DB-VAE latent space are separated by the value of track records of the clinical trial sponsor. If we measure the track record T by the number of phase 1 trials previously completed by the clinical trial sponsor, we find that the drugs in the top-left cluster (blue) have a normalized value of $T \geq 1$ while those in the bottom-right cluster (orange) have $T < 1$. The same separation holds if we use another different measure of sponsor track records (e.g., the number of phase 2 or phase 3 trials instead of phase 1). We conclude that the encoder of DB-VAE distinguishes drugs developed by sponsors with large track records, typically

large multinational pharmaceutical companies, from those with limited track records, typically small biotech companies and academic medical centers.

6.4.4 Improving financial efficiency of drug development

As shown in Section 6.4.1, applying debiasing achieves higher TP and lower TN than the un-debiased counterpart. This tradeoff between higher TP and lower TN leads to an overall improvement of the financial efficiency of drug development, since the revenue generated by correctly predicting a successful drug (the true positive) far outweighs the costs saved by correctly predicting a failed drug (the true negative). We use a simple financial model to illustrate this. Suppose the drug developer has completed a phase 2 clinical trial for a drug candidate and must decide whether or not to conduct a large-scale phase 3 clinical trial. We assume that the phase 3 trial costs \$100 million and takes 5 years to complete. If approved by the FDA, the drug will generate an annual profit of \$2 billion over a 10-year period of market exclusivity. Assuming 10% cost of capital per annum for cash flows of an approved drug and 15% cost of capital for cash flows of phase 3 trial, the net present value (NPV) of an approved drug is $\text{NPV}_1 = \$6$ billion, while that of a drug which fails phase 3 trial is $\text{NPV}_0 = -\$100$ million. The assumed values of the costs of capital are taken from the finance literature [6] and the calculation details are presented in Section D.1. The drug developer uses a machine learning model (with TP and TN) to forecast the approval outcome of the drug candidate from its phase 2 results. The financial value V of the machine learning model is given by

$$V = \text{PoS} \cdot \text{TP} \cdot \text{NPV}_1 + (1 - \text{PoS}) \cdot (1 - \text{TN}) \cdot \text{NPV}_0 \tag{6.4}$$

We evaluate the financial values in six major therapeutic areas with the most recent PoS estimates by Project ALPHA in Q2 2021 [3]. The results are summarized in Table 6.5. Compared with the un-debiased baseline (column 3), applying debiasing to the prediction model generates additional financial values ranging from \$763 to \$1365 million in all six major therapeutic areas. This illustrates the critical role of

debiasing in improving financial efficiency and reducing the financial risks of late-stage drug development.

For the two DB-Label models which debias the outcome labels (columns 5 and 6 of Table 6.5), the additional financial value of DB-Latent is most significant in oncology ($211 million) and cardiovascular diseases ($ 170 million), while negative in metabolic (−$41 million) and infectious diseases (−$42 million). The differences in financial values for different therapeutic areas are correlated with their average debiasing resampling weights $W(z(x)|X_{test})$, shown in Table 6.6. We find that oncology and cardiovascular drugs have the lowest $W(z(x)|X_{test})$ (i.e., they are overrepresented in the latent space distribution). Since debiasing the latent space distribution leads to greatest improvements in TP for overrepresented drugs (Figure 6-3), it makes sense that the increase in financial value is also highest in these two therapeutic areas. On the other hand, metabolic and infectious disease drugs are underrepresented (i.e., they have higher $W(z(x)|X_{test})$), and we do not observe the increase in financial value for these drugs by using DB-Latent.

Table 6.5: Net present value (NPV) to drug developer by using debiased models to predict drug development outcomes.

Therapeutic Area	PoS (%)	No-DB-Label, No-DB-Latent	No-DB-Label, DB-Latent	DB-Label, No-DB-Latent	DB-Label, DB-Latent	Financial Value of Debiasing
Oncology	29.7	238	314	790	1001	763
Cardiovascular	47.2	151	340	1298	1468	1317
Central Nervous System	36.5	209	273	1326	1376	1167
Autoimmune/Inflammation	47.2	585	524	1479	1509	924
Metabolic	46.3	524	483	1876	1835	1311
Infectious Disease	49.6	576	626	1983	1941	1365

Financial value of debiasing (last column) is estimated by the difference between the model which applies debiasing to both outcome labels and latent features (column 3) and the one with no debiasing (column 6). NPV is measured in million $. Both DB-Latent models use $\alpha = 10^{-7}$.

Table 6.6: Average debiasing resampling weights $W(z(x)|X_{test})$ of drugs in each therapeutic area.

Therapeutic Area	No-DB-Label, No-DB-Latent	No-DB-Label, DB-Latent	DB-Label, No-DB-Latent	DB-Label, DB-Latent
Oncology	2.94	2.83	2.93	2.88
Cardiovascular	3.61	3.43	3.56	3.21
Central Nervous System	4.04	3.87	3.81	3.91
Autoimmune/Inflammation	4.13	3.83	3.98	3.96
Metabolic	4.29	3.76	4.58	4.14
Infectious Disease	3.83	3.72	3.84	3.91

Numerical values are measured in units of 0.01%. Oncology and cardiovascular drugs have the lowest $W(z(x)|X_{test})$ for all models.

6.5 Discussion

6.5.1 Implications of debiasing drug approval prediction

Debiasing improves the financial efficiency of drug development and the overall prediction performance (measured by F_1 score) by achieving a higher true positive (TP) rate with a lower true negative (TN) rate. This tradeoff between higher TP and lower TN has important implications for pharmacovigilance. The debiased model is more likely to correctly identify drug candidates with are safe and effective (higher TP) than the un-debiased counterparts. Meanwhile, it may predict a high probability of success for drug candidates which have adverse effects (lower TN). Due to the potential risks of adverse effects, the predictions of the debiased model must be used with caution by the drug developer. To address the drug safety concerns, future works may use the debiased model to predict whether a drug will exhibit adverse effects on a particular patient population.

Also, it is somewhat surprising that DB-VAE achieves greater improvements for drugs which are overrepresented in the latent space distribution with lower debiasing resampling weights $W(z(x)|X_{test})$. This is contrary to the findings of [42] for image classification, where DB-VAE improved the prediction for underrepresented minority subgroups. One possible explanation is that the overrepresented drugs correspond to the "me too" drugs [105] or repurposed drugs which have similar drug properties as previously approved drugs. However, prior approval in one indication does not necessarily lead to a higher probability of success in other indications, which makes the approval outcomes of overrepresented drugs more difficult to predict.

6.5.2 Limitations

Our analysis has several limitations which need to be addressed in future works. First, there are sources of bias in clinical trial development which are important in practice but not addressed by our paper. One example is the patient selection bias against certain demographic features such as race, gender, and socioeconomic status. A ma-

jor difficulty in performing statistical analysis on the bias in patient demographics is the significant under-reporting of the relevant information. As of December 2, 2021, the Informa dataset contains 374,460 clinical trials in all phases of clinical development. Among these, only 1,589 trials (0.4%) recorded "white" or "Caucasian" in the "Patient Population" entry, 1,149 (0.3%) recorded "black" or "African American", and 82 (0.02%) recorded "Latino". Future works should use natural language processing techniques to extract patient demographic information and apply debiasing on the patient demographic features.

In addition, we observe that debiasing the latent space (DB-Latent) improves the true positive rate for drugs which are overrepresented in the feature space (i.e., have similar features). However, our hypothesis that some of the overrepresented drugs are "me too" drugs needs to be confirmed, since the Informa dataset does not explicitly label the "me too" drugs. A potential solution is to use the drug novelty metric proposed in [124] based on the Tanimoto distance between the chemical structures of two drugs. Since the Tanimoto distance applies to small molecules but not necessarily to large biologics, future work needs to generalize the drug novelty metrics to all therapeutics (including combination therapies) and test whether debiasing improves the prediction for "me too" drugs.

Finally, the goal of our work is to illustrate the benefits of applying debiasing on a machine learning model with fixed structure. Despite its improved prediction performance, the debiased model has Type I error (false positive) rate 12% and Type II error (false negative) rate 40%, which should be further reduced by optimizing the model design. In our work, the debiasing and prediction tasks are simultaneously achieved with one neural network. While this is an efficient design, the capacity of the encoder network of DB-VAE may be constrained by performing the dual tasks of reconstructing the input features and predicting the drug approval outcome. A natural extension is to implement debiasing resampling with other prediction models that may be better suited to learn tabular data.

6.6 Conclusion

By instantiating the DB-VAE to our purposes, we simultaneously identify and mitigate the bias from the imbalance of approval outcomes and the over/underrepresentation in drug feature space. We find that debiasing the imbalance of drug approval outcomes results in major improvements in the true positive rate and F_1 score for all drugs, and debiasing the imbalance in feature space improves the true positive rate for overrepresented drugs such as oncology and cardiovascular drugs. The debiased machine learning model predicts safe and effective drugs more accurately and generates financial value for the drug developer in six major therapeutic areas. Future work should address the patient selection bias based on demographic features, incorporate measures of drug novelty as an input feature, and optimize the design of the debiased model to further improve the prediction performance.

Chapter 7

Predicting the Duration of Clinical Trials

The long duration of clinical trials lowers the financial value of novel drug development and delays patients' access to potentially effective treatments. Accurate prediction of the trial duration facilitates more efficient allocation of capital and resources for pharmaceutical companies to operate the clinical trial. We apply traditional survival analysis methods and machine learning models to predict the trial duration using the largest dataset in this domain. We find that the gradient boosting trees achieve the optimal prediction performance and identify key factors which correlate with trial duration. Our methodology and results may help clinical researchers optimize the trial design for expedited testing. [1]

7.1 Introduction

Despite groundbreaking advances in biomedicine, there remains a significant funding gap in financing translational biomedical research from preclinical animal studies to phase 3 clinical trials, a phenomenon known as the "valley of death" of novel drug development [7]. Three institutional challenges to bridging the funding gap are the low probability of success [2], significant capital investments [4], and long duration of

[1] Joint work with Joonhyuk Cho, Chi Heem Wong, and Andrew W. Lo.

clinical trials [5]. While the low probability of success can be effectively remedied via the "multiple shots on goal" approach of parallel drug discovery [9], the long duration of clinical trials is often necessary to recruit sufficiently many patients to demonstrate the safety and efficacy of the drug candidate with a target significance level and power. Martin *et al.* [5] analyzed more than 17,000 trials and found that the median duration of phase 2 trials increased from 33 months in 2008 to 40 months in 2015, while the median duration of phase 3 trials increased from 33 to 39 months during the same period. For the pharmaceutical companies, the long duration decreases the financial value of novel drug development, since it discounts future revenues of drug sales (if the drug is approved) and increases the capital needed to operate the clinical testing sites and perform interim data analysis. For the patients, the long duration prevents potentially effective therapies from reaching those who are direly in need for cure.

To address the challenge of long trial duration, novel trial designs have been proposed and implemented to expedite clinical testing without sacrificing its statistical significance and power. Master protocols, including basket, umbrella, and platform trials, allow concurrent clinical testing of multiple drug candidates or diseases, often with the shared control arm [68], [125]. For diseases with no effective treatments, patients may be willing to accept higher risks of adverse effects (higher Type I error) in exchange for expedited approvals of effective treatments (lower Type II error). Novel trial designs based on Bayesian decision analysis strike the optimal balance between the Type I and II errors for different diseases based on disease severity [13]–[15], [17], [27], [73]. For certain infectious diseases, human challenge trials (HCT) are employed to expedite the clinical testing of vaccine candidates. A recent simulation analysis [34] revealed that the timely initiation and expedited execution of HCT are critical in preventing a large number of infected cases and deaths for COVID-19.

While these unconventional trial designs are employed under special circumstances, it is also important to systematically analyze the common factors which impact the duration for all trials and accurately predict the duration of future trials with these factors. An accurate prediction of trial duration not only facilitates more efficient allocation of capital and resources by pharmaceutical companies to operate clinical

testing, but may also help clinical researchers shorten the trial duration by optimizing the trial design. To the best of our knowledge, our work is the first to apply both traditional statistical methods and novel machine learning models of survival analysis to predict clinical trial duration using the largest dataset in this domain.

7.2 Literature Review

There is a rich literature in estimating clinical trial duration due to its practical importance to the pharmaceutical companies. For clinical trials whose primary outcome is largely uncensored (e.g., COVID-19 infection within 14 days after vaccination), the trial duration is typically estimated using the expected number of patients needed to demonstrate the target significance level and power of the trial, as well as the expected patient enrollment rate [126]. For event-based trials whose primary outcome is censored (e.g., long-term survival or disease progression), accurate estimation of the survival function (i.e., the probability distribution of time-to-event) is also essential to predicting the time of interim analysis and trial duration [127]. Early works in this domain use parametric stochastic processes (e.g., Poisson process and its extensions) to model patient enrollment [128], [129]. Bayesian techniques are commonly used to update the probability distribution of trial duration with the observed time-to-event of enrolled patients [130]–[132]. These parametric statistical models impose strong assumptions on the distribution of patient enrollment and time-to-event. As a result, the prediction accuracy is poor if the model is misspecified [126]. In recent years, with the rapidly growing data of clinical trials, machine learning models are increasingly used to predict patient enrollment. Liu *et al.* [133] train machine learning models to predict the time of 50%, 90% and total enrollment using trial features such as disease indication, trial phase, sponsor, and location. These "bottom-up" approaches in the literature focus on predicting the enrollment rate. However, they are either tailored to model specific types of trials (e.g., immuno-oncology trials as in [134]) or do not use sufficient empirical data to validate their prediction [135].

Our work contributes to the literature in two main aspects. First, in contrast

to the "bottom-up" approach to predict patient enrollment per period, we take the "top-down" approach and directly predict trial duration from a wide variety of trial features using the Citeline Informa database with more than 86,000 trials [22]. In addition, we compare the prediction performance of traditional statistical methods and machine learning models and identify the key factors which correlate with duration. To handle ongoing trials whose durations are right-censored, we apply the statistical and machine learning models in the domain of survival analysis. The models used in our analysis are systematically reviewed by [43]. Several previous works [21], [24], [28] trained machine learning models with the Informa database to predict novel drug development outcomes and provided the methodology of data query for our work.

7.3 Data and Methods

7.3.1 Data Query and Preprocessing

We query historical clinical trial data from the Citeline Informa database [22], one of the largest datasets in this domain. We use the same data query procedure as in [21]. Detailed descriptions of the trial features are summarized in Table E.1. The trial features are either categorical or continuous. For multi-labeled categorical features (e.g., the drug developer may conduct clinical trials in different countries for a drug) with k categories, we apply one-hot encoding and use the k binary child features in our analysis. In contrast to previous works in this domain, the time series of monthly patient enrollment is not included in the trial features in our analysis as this information is not systematically curated by the Informa database. Instead, the number of target patient accrual (which is specified before the trial begins) and actual patient accrual (which is recorded after the trial ends) are both included in the trial features.

To preprocess the raw data for the machine learning models, we first exclude trials with unknown start dates, since we cannot reliably compute their durations. We also exclude trials whose end year occurred before 2000, due to the significant proportion

of missing features in these trials. For trials with known start dates but unknown end dates, we exclude the trials whose outcome status is known (i.e., the trial has ended) or whose start year occurred before 2017. For trials with unknown outcome status and start year no earlier than 2017, we assume that these trials are ongoing and consider their duration to be right-censored since their end dates will occur in the future of our analysis. The preprocessed data consists of 86,838 clinical trials which test 12,454 drugs and 330 disease indications, by far the largest reported in the literature. The summary statistics is shown in Table 7.1.

Table 7.1: Summary statistics of clinical trial duration (years) in Informa dataset.

Phase	Trials	Drugs	Duration Mean	Duration SD	Duration 25%Qt.	Duration Median	Duration 75%Qt.
1	20260	7782	2.3	2.1	0.7	1.7	3.2
1/2	7455	3246	3.6	2.5	1.8	3.0	4.8
2	36066	6486	3.4	2.5	1.7	2.8	4.5
2/3	1905	1122	3.4	2.5	1.6	2.8	4.5
3	21152	3797	3.4	2.5	1.7	2.7	4.3
Total	86838	12454	3.2	2.5	1.4	2.6	4.2

Due to different standards of post-study reporting of clinical trial results (especially before the 2007 FDA Amendments Act), there is considerable missingness in certain clinical trial features which we impute via median imputation. We compute the duration of the trial (measured in years) from its start and end dates, assuming 365 calendar days in a year. To ensure convergence of the statistical and machine learning models trained on our dataset, we remove the trial features whose variance are below 0.05.

7.3.2 Traditional Survival Analysis

We introduce the notation used in the rest of the paper. Let $T_i > 0$ denote the duration of the i-th clinical trial in the dataset and X_i denote its d-dimensional

feature vector. The survival function $S(t) = \mathbb{P}(T > t)$ is the probability that the duration is longer than t. Under mild regularity conditions on $S(t)$, a mathematically equivalent and useful way to characterize the survival function $S(t)$ is through its hazard function $h(t)$ defined by

$$h(t) = \lim_{\Delta t \to 0} \frac{\mathbb{P}(t \leq T < t + \Delta t | T > t)}{\Delta t} = -\frac{S'(t)}{S(t)} \tag{7.1}$$

Note that $h(t) \geq 0$ since $S(t)$ monotonically decreases from 1 to 0 as $t \to \infty$. Since the duration of an currently ongoing trial is right-censored (i.e., observed in the future of our analysis), we use the binary variable δ_i to denote whether the duration T_i of the i-th clinical trial is right-censored ($\delta_i = 1$) or observed without censoring ($\delta_i = 0$).

We use non-parametric (Kaplan-Meier), semi-parametric (Cox regression), and parametric (Weibull accelerated failure time, AFT) statistical models of survival analysis to estimate the survival function $S(t)$ and predict the trial duration by the median time $t_{1/2}$ of $S(t)$, i.e. $S(t_{1/2}) = 1/2$. We use the models implemented in the lifelines library of Python 3.8 [136].

7.3.3 Machine Learning Models

Tree-based Algorithms

Decision tree is a commonly used machine learning model which is simple to train and easy to interpret [137]. During the training stage, each tree node is split into subsequent child nodes by maximizing the homogeneity of data samples within each child node. Common metrics of homogeneity include mean-squared error (for regression) and entropy (for classification). We use the decision tree models for survival analysis implemented in the scikit-survival v0.17.1 library of Python 3.8 [138].

Survival Tree

To train the survival tree on the training dataset, each intermediary tree node is split into its child nodes by maximizing the value of the log-rank test [139]. For

each terminal leaf node n, a Kaplan-Meier survival function $S_n(t)$ is computed with the durations of clinical trials in this node. To evaluate the prediction performance during testing stage, the survival function $S_n(t)$ is used to predict the duration of a new clinical trial which is assigned to node n by applying the partition rules on its features X_i. To prevent overfitting the training dataset, we choose the regularization hyperparameters: maximum depth of the tree $d_{max} = 5$, minimum samples of each split $n_{split} = 15$, and minimum sample per terminal leaf $n_{leaf} = 15$ after tuning the hyperparameter values.

Random Survival Forest

Although the survival tree is simple to train and highly interpretable, its structure is also unstable and sensitive to the distribution of data samples in the training dataset. The random survival forest effectively reduces the variances of the survival tree model via the "bagging" approach, i.e., by training multiple survival trees, each with its training data bootstrapped from the original dataset [140]. To predict the duration of a clinical trial in the testing dataset, the predictions of all survival trees in the forest are averaged. We choose the hyperparameter values: number of survival trees $n_{tree} = 100$, minimum samples of each split $n_{split} = 100$, minimum sample per terminal leaf $n_{leaf} = 100$, and make each split during training by the splitting ratio $r_{split} = 30\%$ of total number of features.

Gradient Boosting Survival Analysis

Gradient boosting tree (GBT) improves decision trees via the technique of boosting [141], i.e., iteratively reducing the residual prediction error of previously trained trees by adding a new tree to the ensemble. Gradient boosting survival trees apply the same technique to survival trees [142]. We train a gradient boosting survival tree which minimizes the partial likelihood loss of Cox's proportional hazards model:

$$L_{GBST} = \sum_{i=1}^{n} \delta_i \left[f\left(X_i\right) - \log\left(\sum_{j \in R_i} \exp(f\left(X_j\right)) \right) \right] \tag{7.2}$$

where $f(X_i)$ denotes the hazard function which is the weighted average of the outputs from all decision trees in the gradient boosting ensemble. We choose the hyperparameter values: number of estimators $n_{est} = 80$, maximum depth of each tree $d_{max} = 5$, minimum sample per terminal leaf $n_{leaf} = 100$, and the learning rate $\alpha = 0.1$.

Neural network-based Algorithms

Development of deep learning algorithms based on neural networks have largely led to the artificial intelligence revolution over the past decade and outperform traditional machine learning models in domains such as computer vision, natural language processing, and reinforcement learning [143]–[145]. In recent years, neural networks have also been applied to survival analysis [43]. We train two neural network models implemented in the pycox library of Python 3.8.

DeepSurv

DeepSurv is the nonlinear generalization of the traditional Cox proportional hazard model [146]. Instead of using a linear function $\beta^T X_i$ in the exponent of the hazard function, DeepSurv uses a nonlinear function $h_\theta(X_i)$ which takes the functional form of a feedforward neural network parameterized by θ. This generalization greatly increases the model's capacity to model nonlinear impact of trial features X_i on the trial duration. We train a neural network $h_\theta(x)$ with two hidden layers of dimension 200 and 100, respectively. We choose ReLU as the activation function [147] and optimize the model parameters via the Adam algorithm [148] with batch size of 256, number of training epochs 300, dropout rate of 0.2, and learning rate $\alpha = 5 \times 10^{-4}$.

Neural Multi-Task Logistic Regression

Similar to DeepCox, the neural multi-task logistic regression is a nonlinear generalization of the traditional multi-task logistic regression (MTLR) model for survival analysis [149]. Traditional MTLR first partitions future time into N intervals $0 = t_0 < t_1 < \ ... < t_{N-1} < t_N = T_{max}$ and uses a logistic regression on trial features X_i to predict the probability p_n that the trial duration T_i is in the n-th interval

$T_i \in [t_{n-1}, t_n)$. Since it performs a separate logistic regression for each time interval, it bypass the strong proportional hazard assumption of the Cox model. The neural MTLR replaces the linear logit in logistic regression with a nonlinear feedforward neural network $h_i(x)$ to compute p_n. For a given partition of time intervals, the survival function $S(t)$ is piece-wise constant in each interval $[t_{n-1}, t_n)$ and is given by $S(t) = \sum_{n=1}^{N} p_n \mathrm{I}\{t < t_n\}$. This becomes a good approximation of the true survival function when the interval lengths $\Delta t = t_n - t_{n-1}$ are chosen to be sufficiently small. We choose the number of partitions $N = 100$ time intervals. The other model hyperparameters are the same as DeepCox.

Survival Support Vector Machine

Survival support vector machine (SSVM) is a ranking algorithm which predicts the relative order of durations (i.e., the binary outcome $\mathrm{I}\{T_i > T_j\}$) for a pair of clinical trials i and j rather than predicting their actual durations T_i and T_j individually [150]. If both T_i and T_j are right-censored, their ranking cannot be determined. However, if at least one of T_i and T_j is observed (e.g., $\delta_j = 0$), their ranking can be determined if one of the two conditions holds: (1) T_i is also observed or (2) T_i is censored but $T_i > T_j$. Formally, the pairs of clinical trials whose durations may be ranked are:

$$P = \{(i, j) \mid T_i > T_j, \quad \delta_j = 0, \quad 1 \leq i, j \leq n\} \tag{7.3}$$

and we train SSVM by minimizing the loss function:

$$L_{SSVM} = \min_{w} \frac{\gamma}{2}|w|^2 + \sum_{i,j \in P} \max\{0, 1 - w^T(x_i - x_j)\} \tag{7.4}$$

where w are the model parameters learned from the data and γ is the L_2 regularization parameter (which we set to $\gamma = 0.001$). The prediction performance is measured by Harrell's concordance index [151], which is the ratio of the number of correctly ranked pairs of trial durations to all comparable pairs P. We use the SSVM model implemented in the pysurvival package of Python 3.8 [152].

Metric of prediction performance

We measure the prediction accuracy using the concordance index (c-index) [151] and compare the accuracy of traditional survival models and machine learning models. We also compare the prediction performance using the original trial features and using the weight of evidence (WoE) encoding of categorical features to gauge the effect of data preprocessing. We use 5-fold cross validation split the training and testing datasets and find the best hyperparameters for each model. The performance of each model is calculated by averaging the c-indices of five independent splits of training and testing data.

7.4 Results

7.4.1 Non-parametric analysis

The summary statistics of clinical trial duration of each clinical phase is provided in Table 7.1 and the corresponding Kaplan-Meier survival function $S(t)$ of each phase is shown in Figure 7-1. We observe that the duration of phase 1 trials is the shortest (with an average of 2.3 years and median 1.7 years) while that of phase 1/2 trial is the longest (with an average of 3.6 years and median 3.0 years). The distributions of durations for phase 2, phase 3 and phase 2/3 trial do not show significant difference. This agrees with our expectation since phase 1 trials mainly tests safety and side effects on a small group of patients while phase 1/2 trials simultaneously test for safety, side effects, as well as treatment efficacy.

7.4.2 Prediction performance

The prediction performance (measured by the c-index) of statistical and machine learning models is summarized in Table 7.2. Among the different models, gradient boosting survival tree model shows the highest c-index in datasets using the original trial features (0.713) and using the WoE-encoded features (0.703). Random survival forest and DeepSurv have slightly lower c-indices but comparable to gradient boosting

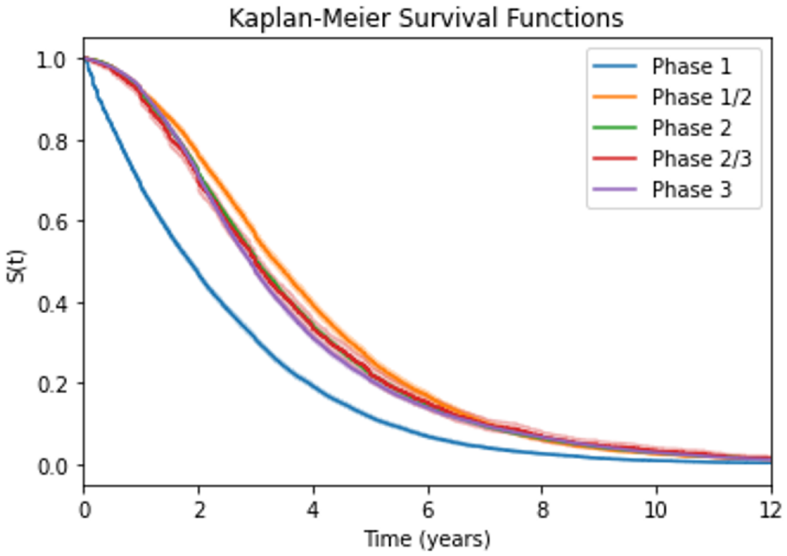


Figure 7-1: Kaplan-Meier survival functions of trial durations for each clinical phase.

trees. This is consistent with the finding of [21] that ensemble models based on decision trees (random forest and gradient boosting tree) achieve the best prediction performance for novel drug development outcomes. On the other hand, the Survival Tree model trained with original trial features shows the lowest concordance index (0.666), possibly because decision trees tend to overfit the training set.

Table 7.2 also shows that models trained with the original clinical trial features consistently overperform the corresponding models trained with WoE-encoded features in all cases except the survival tree. WoE encoding is a useful feature extraction technique to reduce the number of feature dimensions. However, recent studies in the machine learning literature find that ensemble models (e.g., random forest and gradient boosting trees) and neural networks (e.g., DeepSurv, MTLR) with sufficiently large model capacity can automatically extract low-dimensional features from the original high-dimensional features in the training stage [153], [154]. In this case, applying feature engineering such as WoE encoding may weaken the model prediction performance, as observed in our models for trial duration prediction.

Table 7.2: Prediction performance (measured by the c-index) of statistical and machine learning models.

Model	Data Preprocessing	c-Index (Mean)	c-Index (SE)
Cox Regression	Original	0.683	0.001
Weibull AFT	Original	0.684	0.001
Survival Tree	Original	0.666	0.007
	WoE	0.676	0.001
Random Forest	Original	0.701	0.001
	WoE	0.695	0.001
Gradient Boosting Trees	Original	0.713	0.001
	WoE	0.703	0.001
DeepSurv	Original	0.704	0.001
	WoE	0.692	$<$0.001
Neural MTLR	Original	0.683	0.004
	WoE	0.662	0.008
Survival SVM	Original	0.679	0.001
	WoE	0.677	0.001

7.4.3 Feature Importance

Pearson's Correlation Coefficient

As a direct measure of feature importance in predicting the trial duration, we compute Pearson's correlation coefficient ρ between each trial feature and duration. The top 10 features with the largest magnitudes of ρ are listed in Table 7.3. We find that whether a drug treats an oncology indication has the highest $\rho = 0.27$ with trial duration, which is expected due to the long follow-up period to measure the effect of the treatment on long-term survival. The next four features correspond to the types of trial sponsor, which makes sense since government and academic medical centers tend to sponsor trials with smaller size and for rare diseases while the pharmaceutical companies have the incentive and resources to sponsor large-size trials for common

diseases. In addition, we find that whether a drug treats a rare disease has $\rho = 0.13$ with trial duration, which reflects the difficulty of conducting clinical trials for rare diseases despite the smaller patient size.

Permutation Importance

Permutation importance is a generic method to measure the importance of each feature on the prediction performance of any machine learning model [155]. For a given feature, the permutation importance is defined as the decrease of prediction performance if the values of this feature are randomly permuted across the data samples in the testing dataset while the values of all other features remain the same.

Table 7.3 shows the permutation importance of the top 10 trial features with largest magnitudes of correlation ρ with trial duration. Despite the differences in permutation importance assigned by different machine learning models, the top 5 features with highest ρ also have the highest permutation importance overall, which confirms the consistency of the feature importance analysis.

Table 7.3: Top 10 features with largest magnitudes of Pearson's correlation to trial duration and permutation importance.

Feature	Pearson's ρ	DeepSurv	Neural MTLR	Survival Tree	Random Forest	Gradient Boosting Tree	Survival SVM
Indication group: Anticancer products	0.272	0.046	0.038	0.082	0.061	0.057	0.064
Sponsor type: Industry, All other pharma	−0.202	0.010	0.003	0.017	0.011	0.008	0.005
Sponsor type: Academic	0.197	0.019	0.008	0.018	0.023	0.018	0.010
Sponsor Type: Government	0.189	0.008	0.006	0.007	0.006	0.006	0.007
Sponsor type: Cooperative group	0.184	0.006	0.004	0.001	0.003	0.004	0.005
Trial Phase: Phase 1	−0.150	0.008	−0.010	0.025	0.013	0.015	0.001
Medium: Solution	0.138	0.006	0.009	0.000	0.001	0.000	0.001
Drug delivery route: Injectable	0.137	0.006	0.001	0.003	0.003	0.003	0.001
Indication group: Rare diseases	0.128	0.005	0.003	0.000	0.002	0.003	0.001
Indication group: Alimentary / Metabolic	−0.125	0.003	0.001	0.000	0.001	0.001	0.001

Survival Tree	Random Forest	Gradient Boosting Tree	DeepSurv	Neural MTLR	Survival SVM
Indication Group: Anticancer Products	Indication Group: Anticancer Products	Indication Group: Anticancer Products	Indication Group: Anticancer Products	Indication Group: Anticancer Products	Indication Group: Anticancer Products
Trial Phase: Phase 1	Sponsor Type: Academic	Target Accrual	Sponsor Type: Academic	Medium: Solution	Trial Phase: Phase 3
Sponsor Type: Academic	Target Accrual	Percentage Accrual	Percentage Accrual	Sponsor Type: Academic	Trial Phase: Phase 2
Sponsor Type: Industry, All Other Pharma	Trial Phase: Phase 1	Sponsor Type: Academic	Continent: North America	Trial Phase: Phase 2	Sponsor Type: Academic
Trial Phase: Phase 2	Sponsor Type: Industry, All Other Pharma	Trial Phase: Phase 1	Trial Phase: Phase 3	Percentage Accrual	Sponsor Type: Government
Target Accrual	Percentage Accrual	Sponsor Type: Industry, All Other Pharma	Trial Phase: Phase 2	Continent: North America	Continent: Europe
Continent: Asia	Sponsor Type: Government	Sponsor Type: Government	Target Accrual	Sponsor Type: Government	Sponsor Type: Cooperative Group
Sponsor Type: Government	Continent: Asia	Trial Outcome: Completion	Trial Outcome: Completion	Trial Phase: Phase 3	Sponsor Type: Industry, All Other Pharma
Route: Injectable	Trial Phase: Phase 3	Continent: North America	Sponsor Type: Industry, All Other Pharma	Target Accrual	Continent: North America
Trial Phase: Phase 3	Trial Phase: Phase 2	Trial Phase: Phase 2	Sponsor Type: Government	Continent: Europe	Medium: Solution

Figure 7-2: Top 10 features with highest permutation importance for each machine learning model.

To analyze the permutation feature importance of individual models, we list the top 10 most important features of each model in Figure 7-2. The features that are more commonly identified by different machine learning models are shown in darker colors. Despite their differences in prediction performance, the machine learning models consistently identify the same set of trial features with highest impact on prediction performance. Interestingly, all models identify whether the drug treats an oncology indication as the most important feature. In addition, the trial sponsor type (academic or government medical center) and whether the trial is in phase 2 are also identified by all models. The consistency across different machine learning models confirms the impact of these features on trial duration.

In addition, we find that two trial features (target patient accrual t_{acc} and actual patient accrual as a percentage of target accrual p_{acc}) have relatively low Pearson's correlation with trial duration but are identified as the top 10 most important features by the majority of machine learning models (Figure 7-2). This illustrates the power of machine learning models to capture the nonlinear interactions between trial features which, in this case, is the product of $t_{acc}p_{acc}$ and is equal to the total number of accrual patients in the trial.

7.5 Discussion

By applying different survival analysis models to predict the clinical trial duration with the largest dataset in this domain, we systematically identify several key factors which influence the trial duration. The most important factor (measured by both Pearson's correlation and permutation importance) is whether the drug treats an oncology indication. The long duration of oncology trials is partly due to the necessity for post-treatment follow up with the trial participants over an extended period in order to measure the trial endpoints such as long-term patient survival and disease progression. Though the extended follow-up period is required, the overall duration of oncology trials can still be shortened through the wider application of novel trial designs such as platform trials [68] and Bayesian trials [14] which can be tailored for

different diseases.

In addition, clinical trials conducted by academic medical centers and government agencies are significantly longer than those conducted by the pharmaceutical companies. This discrepancy calls for greater public-private partnership in novel drug development, especially for rare diseases which do not generate large revenues for the pharmaceutical companies. If the pharmaceutical companies are given higher incentive to develop drugs for rare diseases (e.g., in the form of priority review voucher), the duration of these trials may be significantly shortened for the benefits of the patients direly in need.

7.6 Conclusion

We apply statistical and machine learning models to predict the duration of clinical trials in the largest dataset of this domain. We find that gradient boosting trees achieve the best prediction performance. Key factors which influence trial duration include whether the drug treats an oncology indication, the type of clinical trial sponsor, the clinical trial phase, and the numbers of target and actual patient accrual. Our results call for the wider use of novel trial designs and greater public-private partnership in order to expedite the clinical development for potentially life-saving therapeutics.

Chapter 8

Estimating the Correlation of Clinical Trial Outcomes

The correlations between clinical trial outcomes are the key parameters which significantly influence the financial performance of the biomedical megafund. However, it is difficult to estimate the correlation from historical drug development data due to the complex biomedical factors which induce the correlations. In this chapter, we use both non-parametric and parametric methods in the biostatistics literature to estimate the correlations from historical data. While the non-parametric estimator does not yield statistically significant correlations, the parametric estimator of generalized estimating equations yields positive and statistically significant correlations across all therapeutic areas ranging from 2.0% (central nervous system) to 7.3% (metabolic). Future works should improve the specification of correlation structure to balance the diversity of biomedical factors included and computational feasibility. [1]

8.1 Introduction

In portfolio management, the correlations between assets have significant impact on the volatility of the portfolio's return. In the context of managing a portfolio of

[1]Joint work with Andrew W. Lo. Research assistance of Jack Zelman and Arturo Chavez-Gehrig in the early stages of this project is gratefully acknowledged.

biomedical assets (e.g., drug candidates currently under clinical testing), a recent simulation study [44] shows that the financial performance of the portfolio becomes less attractive when correlation is introduced across different clinical trials and across different phases of the same clinical trial. Higher correlation decreases the Sharpe ratio and increases the probability of wipeout, where all the drug candidates simultaneously fail to proceed to the next clinical stage. In practice, biomedical expertise and active portfolio management are critical to ensuring that the portfolio is well diversified and can generate financial value for the investors [26].

In previous studies of biomedical megafund simulations [11], [26], a panel of external biomedical experts are asked to evaluate the similarity of each pair of drug candidates in the portfolio and qualitatively assign the levels of "high", "medium" and "low" correlations if the approval of one drug candidate will "significantly", "moderately", or "hardly" increase the probability of success (PoS) of the other drug candidate. The qualitative assessments are transformed into a quantitative correlation matrix by assigning numerical values to each correlation level (e.g., $\rho_{high} = 0.7, \rho_{med} = 0.4, \rho_{low} = 0.1$) and averaging across the estimates of all experts. Since the resulting correlation matrix $\tilde{\Sigma}$ is symmetric but not guaranteed to be positive definite, $\tilde{\Sigma}$ is transformed to the closest symmetric and positive-definite matrix $\hat{\Sigma}$ via a convex optimization method [62]. In practice, one often finds that $\hat{\Sigma}$ and $\tilde{\Sigma}$ are close to each other so that the qualitative assessments of domain experts are well preserved in the financial simulations [26]. The advantage of this approach is that the correlation estimates are biologically motivated. The disadvantage is that the expert assessments are subjective and may be systematically biased. So far, no data-driven methods have been proposed to estimate the correlation from the historical data of clinical trial outcomes. Our work addresses this open problem by applying rigorous techniques in the biostatisitcs literature to estimate correlations using the largest dataset in this domain.

8.2 Data and Methods

8.2.1 Data

We query the historical drug development data from Citeline Informa dataset [22], the largest dataset in this domain with over 93,000 drugs and 380,000 clinical trials (as of April 6, 2022). The data query and pre-processing methods follow from the time series PoS estimation method proposed by [2] and utilized by Project ALPHA [3]. We query the number of clinical trials initiated in each year from 2004 to 2018 as well as the clinical phase (1, 2, and 3), target disease (e.g., epilepsy, renal cancer, etc.), and outcome (success or failure) of each clinical trial. We do not include the data from 2019 to 2021 since the outcomes of most clinical trials initiated during this period are still unknown. The summary statistics is shown in Table 8.1.

Table 8.1: Summary statistics of annual clinical trial outcomes in Informa dataset.

Trial Start Year	Success Phase 1	Total Phase 1	Success Phase 2	Total Phase 2	Success Phase 3	Total Phase 3
2004	631	1313	463	1250	236	663
2005	755	1487	452	1357	240	664
2006	647	1550	411	1409	301	803
2007	860	2060	454	1649	254	678
2008	733	2207	398	1571	202	597
2009	861	2242	413	1546	221	543
2010	776	2142	368	1472	206	531
2011	865	2212	352	1343	268	629
2012	685	1928	371	1349	182	542
2013	657	1852	296	1172	207	533
2014	790	2159	300	1261	225	508
2015	859	2088	298	1096	210	492
2016	807	1972	313	1104	198	384
2017	764	1961	227	890	132	224
2018	656	1826	197	512	132	151

8.2.2 Notation

We use $Y_i \in \{0, 1\}$ to denote the binary outcome of the drug development outcome for drug i and $X_i \in \mathbb{R}^d$ to denote the associated feature vector of the drug and clinical trial. The features in X_i can be either continuous (e.g., number of patients enrolled in the clinical trial) or categorical (e.g., whether the drug treats an oncology indication). Additional subscripts are introduced when we further stratify the drugs into K clusters (e.g., based on the therapeutic mechanism or the year of clinical trial outcome). For cluster k, let $Y_{i,k}$ denote the outcome of drug i, N_k the number of drugs, and $S_k = \sum_{i=1}^{N_k} Y_{i,k}$ the total number of successful clinical trials, respectively. Let $N = \sum_{k=1}^{K} N_K$ denote the total number of clinical trial outcomes.

8.2.3 Non-parametric Correlation Estimator

In biostatistics, correlated binary variables often arise in two scenarios. First, the treatment and control groups of a clinical trial are randomized on the cluster level and the interactions among the subjects in each cluster cannot be ignored due to possible confounding [156]. Second, multiple observations represent the longitudinal measurements of the same subject (e.g., in a long-term follow-up study [157]) and are naturally correlated as a time series.

There is a rich literature in biostatistics on estimating the intra-class correlation coefficient (ICC) of patient responses (e.g., see [45] and [46] for comparative studies of different ICC estimators). These estimators assume a common correlation model, where the binary outcomes in the same cluster k have the same PoS $\mathbb{P}(Y_{ik} = 1) = \pi$ and Pearson's correlation $\text{Corr}(Y_{ik}, Y_{jk}) = \rho$ where $i, j \in [N_k]$ and $i \neq j$. The outcomes across different clusters are assumed to be independent. Under these assumptions, there are three types of commonly used non-parametric ICC estimators.

Estimator 1: ANOVA estimator

$$\hat{\rho}_A = \frac{\text{MS}_B - \text{MS}_W}{\text{MS}_B + (n_A - 1)\text{MS}_W} \tag{8.1}$$

with auxiliary variables

$$
\begin{aligned}
n_A &= \frac{1}{K-1}\Big(N - \frac{\sum_{k=1}^{K} n_k^2}{N}\Big) \\
\mathrm{MS}_B &= \frac{1}{K-1}\Big(\sum_{k=1}^{K} \frac{S_k^2}{N_k} - \frac{(\sum_{k=1}^{K} S_k)^2}{N}\Big) \\
\mathrm{MS}_W &= \frac{1}{N-K}\Big(\sum_{k=1}^{K} S_k - \sum_{k=1}^{K} \frac{S_k^2}{N_k}\Big)
\end{aligned}
\tag{8.2}
$$

Estimator 2: Fleiss-Cuzick (FC) estimator

$$
\hat{\rho}_{FC} = 1 - \frac{1}{(N-K)\hat{\pi}(1-\hat{\pi})} \sum_{k=1}^{K} \frac{S_k(n_k - S_k)}{n_k} \tag{8.3}
$$

with auxiliary variable $\hat{\pi} = \frac{\sum_{k=1}^{K} S_k}{N}$.

Estimator 3: Pearson estimator

$$
\hat{\rho}_P = \frac{1}{\hat{\mu}(1-\hat{\mu})}\Big(\frac{\sum_{k=1}^{K} S_k(S_k-1)}{\sum_{k=1}^{K} N_k(N_k-1)} - \hat{\mu}^2\Big) \tag{8.4}
$$

with auxiliary variable $\hat{\mu} = \frac{\sum_{k=1}^{K} S_k(N_k-1)}{\sum_{k=1}^{K} N_k(N_k-1)}$.

The main advantage of using these ICC estimators is that their finite-sample standard errors can be derived exactly and do not rely on asymptotic approximations such as the central limit theorem. This is particularly useful in our application since the number of clinical trial outcomes may not be sufficiently large for asymptotic approximations in certain therapeutic areas such as rare diseases [10]. The expressions for standard errors of these ICC estimators are complicated and given in Section F.1. We implement the three ICC estimators in Python 3.8.

8.2.4 Parametric Correlation Estimator

The non-parametric ICC estimators are simple to implement and have exact finite-sample standard errors. However, the assumption of uniform correlation within each cluster may be too simplistic since they do not account for the biological factors which induce different degrees of correlations among the clinical trial outcomes. For instance, one might believe that the clinical trial outcomes of vaccine candidates for COVID-19 are correlated since they target the same disease. In addition, the vaccine candidates which use the same therapeutic mechanism (e.g., messenger RNA technology) may have higher correlation since this therapeutic mechanism might be particularly effective against COVID-19. To quantitatively capture the effect of relevant drug and clinical trial features on the correlation, we need to use parametric estimators.

The parametric estimators of correlation in general fall into two classes. The first is the latent variable models, where we assume there exists an unobserved latent variable W_{ik} and the binary outcomes are generated by $Y_{ik} = \mathrm{I}\{W_{ik} > 0\}$. The advantage of the latent variable model is that the correlations of $W_k = [W_{1k}, ..., W_{N_k k}]$ are often much simpler to specify (e.g., through a linear factor structure) and naturally induce correlations among Y_{ik}. As a result, latent variable models are commonly used in financial simulations for correlated defaults [44], [158], [159]. However, the estimation of model parameters is numerically difficult since we need to compute the likelihood function based on observed outcomes Y_{ik} by integrating over the latent variables W_k. Common numerical integration techniques such as Gauss-Hermite quadrature are not suitable when there are more than one latent factor [160], [161].

A different class of parametric estimators, known as the generalized estimating equation (GEE) and pioneered by the seminal work of [47], is one of the most popular inference methods in biostatistics due to its capacity to model diverse correlation structures and yield consistent estimators for the parameters in the marginal PoS even if the correlation structure is misspecified. The original work [47] focused on estimating the parameters β for the PoS and treating the parameters α for correlation

as nuisance parameters. Subsequent works derived jack-knife estimators of standard errors [162] and extended the original GEE to allow joint estimation of β and α via modified Fisher scoring update [49].

In contrast to the latent variable models, GEE is particularly suited for inference since we only need to specify the functional form of marginal PoS for each Y_{ik} (e.g., as a logit or probit regression on the covariates X_{ik} with regression parameters β) as well as the functional form of the correlation $\rho_{ij}(\alpha)$ between outcomes Y_{ik} and Y_{jk} which depends on the parameters α and the covariates X_{ik} and X_{jk}. Neither the joint distribution of the binary outcomes nor the underlying stochastic process which generates these outcomes (e.g., W_{ik} in the latent variable model) needs to be specified. In the simplest case, we can impose that the binary variables are equi-correlated by setting $\rho_{ij}(\alpha) = \alpha$. More complicated correlation structures motivated by the biomedical domain knowledge are also feasible. For instance, let the covariate X_{i1} denote the target disease of drug candidate i (e.g., pancreatic cancer, COVID-19) and X_{i2} denote its clinical trial phase (1, 2, or 3). A correlation matrix in the linear factor form is $\rho_{ij}(\alpha) = \alpha_0 + \alpha_1 \mathrm{I}\{X_{i1} = X_{j1}\} + \alpha_2 \mathrm{I}\{X_{i2} = X_{j2}\}$ where the parameters α_1, α_2 capture the additional correlation induced by the common disease and clinical phase, while α_0 is the baseline correlation in the absence of any common features.

We use the GEE estimation package *geepack* (version 1.3.3) in the R language which implements the extended version of GEE for joint estimation of PoS and correlation parameters with their associated standard errors [49], [163], [164].

8.3 Results

8.3.1 Non-parametric Correlation Estimator

We apply the non-parametric ICC estimators to the clinical trial outcomes in each therapeutic area. We estimate the correlation for clinical trials in each phase separately since the PoS varies significantly between clinical phases due to the different clinical endpoints and patient enrollment [2]. As a result, the common correlation

assumption of ICC estimators is not satisfied if we jointly estimate the correlation for all phases. We assume that the clinical trials targeting different diseases are independent and form the independent clusters by their target diseases. The estimated values of the ICC for three different estimators are summarized in Table 8.2.

Overall, we find that the ICC is positive in all therapeutic areas and clinical trial phases. In addition, the ICC of phase 3 trials is higher than phase 1 and phase 2. The three ICC estimators generally yield consistent estimation results. However, the value of estimated ICC is below its associated standard error and we do not observe statistically significant correlations between the clinical trial outcomes.

One possible explanation of the large standard errors of the ICC estimators is the large variation in the conditional PoS across different target diseases (e.g., a drug for migraine is more likely to receive FDA approval than a drug for pancreatic cancer). This motivates using the drug and clinical trial features to estimate the conditional PoS and correlation with the GEE approach, as discussed in the next section.

Table 8.2: Intra-class correlation estimates of clincal trial outcomes in the Informa dataset.

Therapeutic Area	Phase	$\hat{\rho}_A$	SE	$\hat{\rho}_{FC}$	SE	$\hat{\rho}_P$	SE
All	1	0.043	0.092	0.042	0.025	0.026	0.031
	2	0.060	0.138	0.059	0.032	0.037	0.041
	3	0.161	0.341	0.160	0.042	0.149	0.065
Oncology	1	0.043	0.100	0.042	0.039	0.029	0.044
	2	0.023	0.071	0.022	0.042	0.015	0.045
	3	0.077	0.196	0.074	0.071	0.065	0.101
Metabolic	1	0.033	0.108	0.029	0.062	0.005	0.053
	2	0.051	0.154	0.046	0.080	0.022	0.097
	3	0.147	0.355	0.133	0.113	0.123	0.172
Cardiovascular	1	0.031	0.112	0.025	0.066	0.010	0.058
	2	0.029	0.094	0.025	0.055	0.017	0.053
	3	0.089	0.242	0.078	0.104	0.069	0.116
Central Nervous System	1	0.049	0.132	0.045	0.064	0.027	0.071
	2	0.069	0.192	0.064	0.088	0.034	0.112
	3	0.099	0.271	0.089	0.112	0.034	0.132
Autoimmune/ Inflammation	1	0.015	0.051	0.013	0.034	0.004	0.025
	2	0.031	0.097	0.029	0.057	0.011	0.047
	3	0.047	0.124	0.043	0.059	0.019	0.053
Infectious Disease	1	0.062	0.183	0.054	0.091	0.025	0.085
	2	0.094	0.242	0.083	0.099	0.040	0.098
	3	0.100	0.258	0.088	0.103	0.050	0.114

8.3.2 GEE Correlation Estimator

Under the GEE approach, we need to specify the functional forms of the marginal PoS and the correlation matrix using the generalized linear model (GLM) specification. We use the target disease and the phase of the clinical trial as regression covariates. Since the target disease and clinical phase are both multi-level categorical variables, we apply one-hot encoding and create the binary vector of child features D_i for target disease and P_i for clinical phase. We denote the covariates by $X_i' = [P_i', D_i']$.

We assume that the marginal PoS of the trial outcome Y_i takes the form

$$\mathbb{P}(Y_i = 1|X_i) = g(X_i'\beta) \tag{8.5}$$

where $g(\cdot) : \mathbb{R} \to (0, 1)$ is the mean link function. We choose $g(\cdot)$ to be the logit or probit function and compare the resulting correlation estimates. In addition, we choose the equi-correlated matrix $\rho_{ij} = \alpha$ for clinical trial outcomes Y_i and Y_j in each cluster formed by the target disease. The estimated values of the GEE correlation in each therapeutic area are summarized in Table 8.3.

We find that the estimated correlations are positive in all therapeutic areas, which is consistent with the biomedical intuition. The correlation estimated using the logit link function is consistently higher than the corresponding value estimated using the probit link function. In addition, all estimated correlations are below 8%, with metabolic drugs having the highest correlation (7.3% using logit and 6.2% using probit) and central nervous system drugs having the lowest correlation (2.0% using logit and 1.6% using probit). In contrast to the ICC estimates, the GEE estimates are all statistically significant with p-values of the Wald test no greater than 0.001.

Table 8.3: GEE correlation estimator of clinical trial outcomes in the Informa dataset.

Therapeutic Area	PoS Link	α	SE	p-value
Oncology	probit	0.017	0.002	<0.001
	logit	0.031	0.003	<0.001
Metabolic	probit	0.062	0.011	<0.001
	logit	0.073	0.013	<0.001
Cardiovascular	probit	0.046	0.010	<0.001
	logit	0.053	0.011	<0.001
Central Nervous System	probit	0.016	0.005	0.001
	logit	0.020	0.006	0.001
Autoimmune/ Inflammation	probit	0.026	0.005	<0.001
	logit	0.032	0.007	<0.001
Infectious Disease	probit [2]	NA	NA	NA
	logit	0.056	0.017	0.001

8.4 Discussion

Our work presents a proof of concept study for using rigorous inference techniques to estimate the correlation between clinical trial outcomes, accounting for the effects of covariates such as target disease and clinical trial phase. Using the GEE approach, we find that the correlation is relatively weak (below 8%) yet statistically significant. The numerical values of the estimated correlations are smaller than the values obtained using biomedical expert estimates in previous megafund simulation studies [11], [26]. Our results may potentially have important implications on the active management of biomedical portfolios since the correlation is a key parameter which influence the volatility of the portfolio return.

An crucial limitation of the GEE approach is the restriction on the maximum size of the independent clusters if we also include the covariates X_i to specify the correla-

[2]GEE with the probit link function does not converge for infectious diseases.

tion matrix (e.g., in the linear factor form $\rho_{ij}(\alpha) = \alpha_0 + \alpha_1 \mathrm{I}\{X_{i1} = X_{j1}\} + \alpha_2 \mathrm{I}\{X_{i2} = X_{j2}\}$). For a cluster k of size n_k, the design structure of the correlation matrix is of order $O(n_k^2)$. In practice, a cluster of size greater than 1000 is computationally infeasible using the *geepack* package in R. This limits the model's capacity to capture the complex correlation structure which is jointly induced by biomedical factors such as therapeutic mechanism, clinical trial sponsor, and target disease. As a result of this computational restriction, future extensions of our work should improve the specification of independent clusters to capture both the diverse correlation structure while maintaining computational feasibility.

8.5 Conclusion

We apply both non-parametric and parametric methods in biostatistics to estimate the correlations between historical clinical trial outcomes in the Informa dataset. While the non-parametric ICC estimator does not yield statistically significant correlations, the parametric GEE approach reveals weak correlations (below 8%) in each therapeutic area which are statistically significant. Our estimates potentially have important implications for the management of biomedical portfolios and may be further improved by more appropriate specifications of independent clusters of clinical trials using the drug and clinical trial features.

Part V

Social and Ethical Aspects of Drug Development

Chapter 9

Success and Challenges of a Disruptive Drug Pricing Strategy

We examine the success and challenges of the disruptive pricing strategy of *abaloparatide*, an osteoporosis drug launched by Radius Health in 2017 at a list price 45% lower than its main competitor. This strategy allowed Radius to gain rapid access to this market and achieve a corresponding growth in patient volume. It now faces two challenges: the perverse incentive of Medicare Part D rebates, and the paradox of Medicare Part D coverage structure that prevents lower list prices from necessarily leading to lower out-of-pocket costs for all patients. Nevertheless, we find that this pricing strategy is sustainable for the drug manufacturer, beneficial for the patient, and may have potential applications in other therapeutic areas.[1]

9.1 Introduction

The healthcare industry in the United States is a complex ecosystem with many different stakeholders. Unlike the universal single-payer healthcare systems of many European countries, the accessibility of prescription drugs in the U.S. is largely determined by contract negotiations between health plans and drug manufacturers about

[1]Joint work with Andrew W. Lo. The previous version of this chapter was published as a research article in *Journal of Investment Management* [29]. Valuable feedbacks from Jayna Cummings are gratefully acknowledged.

formulary placement. These negotiations can sometimes result in higher out-of-pocket costs for the patient, since the current structure of the U.S. healthcare system creates a perverse incentive for many health plans to elicit higher rebates from drug manufacturers in exchange for formulary placement of brand-name drugs, thereby increasing patient out-of-pocket costs.

Despite the landmark reforms of the Affordable Care Act, 28% of adults in the U.S. between the ages of 19 and 64 with full-year health insurance in 2016 were still underinsured, and unable to afford prescribed medication. This is more than twice the corresponding rate in 2003 [165]. The high list price of drugs exerts a direct adverse impact on adherence rates and patient treatment outcomes, especially for patients who have not reached their insurance deductible or who make a coinsurance payment at a fixed percentage of the list price. Similarly, a 2010 study found that prescription drugs with copayment over $50 are nearly five times more likely to be abandoned by the patient at the pharmacy counter than those with no copayment [166]. These high list prices not only impose significant financial burdens on individual patients, but also threaten the public health of general society.

In its healthcare blueprint, "American Patients First," issued in May 2018, the Trump administration identified high list prices and the high out-of-pocket costs of drugs as two major challenges to the U.S. healthcare system [167]. To directly reduce the out-of-pocket costs to patients, the Department of Health and Human Services proposed in January 2019 to replace the rebate-driven system with upfront discounts [168]. However, as the U.S. healthcare system consists of hundreds of widely varying local systems, there are considerable challenges to regulating drug prices at the federal level. In the absence of effective government regulation, the pharmaceutical industry can benefit from fair and responsible pricing strategies that are both financially sustainable for the drug manufacturer and affordable for patients with standard health insurance.

In this case study, we analyze Radius Health's pricing strategy for the drug *abaloparatide*, approved by the U.S. Food Drug Administration (FDA) in 2017 to treat postmenopausal women with osteoporosis at high risk of fracture. With an ini-

tial list price 45% lower than its main competitor, *abaloparatide* managed to achieve rapid market access and significant patient volume growth within twenty months after launch. We discuss the potential of this pricing strategy to become a template for responsible pricing in the pharmaceutical industry.

9.2 Background

The success of Radius Health's pricing strategy is highly specific to the context of anabolic osteoporosis, as are the challenges to it. To understand why its pricing strategy has disrupted the osteoporosis therapeutics market, it is essential to first examine the disease and patient population of osteoporosis, and its market dynamics prior to the launch of *abaloparatide* in 2017.

Women's osteoporosis is a common but largely undertreated disease in the United States. As of 2010, an estimated 8.2 million women above the age of 50 years in the U.S. suffered from osteoporosis [169]. A study in 2014 found that, among a group of 47,171 women over the age of 50 who had experienced an osteoporotic fracture, only 23% of them received treatments for osteoporosis during the first year following the fracture [170]. Because many osteoporosis patients also have other chronic conditions (for example, cardiovascular disease), they tend not to take adequate measures to prevent fractures, especially if the medication incurs a serious financial burden. Once an osteoporosis patient experiences a fracture, however, the treatment is often much more expensive than the preventive medication, and the patient is subject to an increased risk of mortality due to associated complications from the fracture [171].

Currently there are two major categories of treatment for women's osteoporosis: antiresorptive agents and anabolic agents. Antiresorptive agents reduce the rate of bone breakdown, have lower costs, and are administered orally or via injection. Anabolic agents, on the other hand, stimulate the formation of new bones, but have a much higher cost, and require daily self-injection [172].

In 2016, the antiresorptive drug denosumab was the revenue leader in women's osteoporosis, treating 800,000 patients, and capturing all annual growth in patient

volume. In comparison, the anabolic agent segment of the market only treated 48,000 patients in 2016, a market penetration of less than 5%. The patient volume within this segment declined by 45% from 2011 to 2016, largely due to a 250% increase in the list price of *teriparatide*, the only anabolic drug available between 2003 and the launch of *abaloparatide* in 2017. This price increase appears to reflect the lack of long-term commitment of the manufacturer Eli Lilly in women's osteoporosis market. At the time of the increase, Lilly's patent on *teriparatide* was expected to expire by August 2019 [173]. The downturn of the anabolic agent market coupled with looming competition from biosimilar versions of *teriparatide* set the stage for Radius Health's competitive pricing strategy for *abaloparatide*.

9.3 Company History

Radius Health, Inc. is a biopharmaceutical company based in Waltham, Massachusetts. Originally named Nuvios, it was founded in 2003 by a group of academic researchers specializing in endocrinology and bone mineral metabolism with a primary focus on research and development (R&D). In 2005, Radius in-licensed the compound *abaloparatide*, receiving the patent license from the pharmaceutical firm Ipsen to develop a novel anabolic drug for osteoporosis. One year later, Radius in-licensed the compound *elacestrant* from Eisai to initiate the development of a new hormone therapy for late-stage ER+/HER2- breast cancer.

As the clinical program for *abaloparatide* (administered as a subcutaneous injection) progressed to phase 3 in 2011, Radius decided not to partner with a major pharmaceutical company to launch its lead product for the U.S. market, but instead launched *abaloparatide* on its own. In its transition from an R&D firm to a commercial company, Radius completed its initial public offering in June 2014, greatly expanded its sales and marketing departments, and brought new members onto its senior management team with extensive expertise in drug development and commercialization. In April 2017, *abaloparatide* was approved by the FDA to treat postmenopausal women with osteoporosis at high risk of fracture.

Radius Health intends to become a leader in women's health therapeutics in the U.S. Its current pipeline includes a novel transdermal patch formulation of *abaloparatide* (currently in phase 3 clinical trials), *abaloparatide* therapy for men with osteoporosis (phase 3), and the hormone therapy *elacestrant* for late-stage ER+ and HER2- breast cancers (phase 3). The primary focus of its mission is to bring innovative and financially accessible therapies to women with serious health conditions.

9.4 Pricing Strategy of *abaloparatide*

Radius's pricing strategy for its lead product, *abaloparatide*, was developed under the realization that it would need to achieve high sales and rapid gains in market share in order to meet the goals of its stakeholders. Launched in June 2017, *abaloparatide* faced intense competition from the anabolic drug *teriparatide*, manufactured by Eli Lilly since 2002 and covered by the major health plans. Radius believed that a lower list price would help to accelerate the coverage of *abaloparatide* by commercial health plans, Medicaid, and Medicare Part D.

Radius anticipated future competition from biosimilar versions of *teriparatide*. Following the expiration of the patent for *teriparatide* in the second half of 2019 [173], it is projected that biosimilar drugs will enter the anabolic agent market with list prices 15% to 30% lower than that of *teriparatide*. However, *abaloparatide* with a list price 45% lower than that of *teriparatide* should still retain a competitive edge in pricing after the entrance of these biosimilar drugs.

Radius also wanted to enlarge the market for anabolic therapies for osteoporosis. As described earlier, the patient volume of this market had declined by over 45% from 2011 to 2016, largely due to the 250% increase in the list price of *teriparatide* to $35,000 annually. A significantly lower list price would quickly differentiate *abaloparatide* as a product and provide a strong incentive for physicians to prescribe it as the preferred anabolic therapy.

In addition, Radius wanted to demonstrate its commitment to socially responsible drug pricing. Price spikes taken by brand-name drugs, often near the expiration of

their patents, have generated intense public criticism of the ethics of the biopharmaceutical industry. Instead of creating a niche product with a high price and a small patient volume, Radius believed that a lower price for *abaloparatide* would make a state-of-the-art anabolic therapy accessible to a larger patient group.

9.4.1 Success

The first milestone of Radius's pricing strategy was to gain coverage among the commercial and Medicaid segments of the health plan market. Within eight months of approval, *abaloparatide* achieved 92% coverage of patients insured by the commercial health plan market. After twenty months, it had achieved 99% coverage in the commercial market, and 96% in Medicaid, both exceeding the coverage of its main competitor. This milestone was significant, since the commercial and Medicaid segments combined account for 50% of the volume of the anabolic agent market, and 81% of total coverage in the U.S.

By the end of 2018, *abaloparatide* had captured 40% of new-to-brand prescriptions in the anabolic agent market, 31% of new prescriptions, and 27% of total prescriptions, as measured by patients' months on therapy (PMOT). As of 2019, *abaloparatide* is covered at parity or better by five of the seven largest Medicare Part D health plans in the U.S., an increase of 28% in potential anabolic agent market volume, and as of the third quarter of 2019, *abaloparatide* has captured 42% of new prescriptions and 37% of total prescriptions.

It is important to Radius's pricing strategy not only to expand its share of the anabolic agent market, but to expand its patient volume as well. The patient volume, again measured by PMOT, grew on average by 8.5% during each quarter of 2018 over the same quarter in 2017. Presumably, its low list price and high coverage by commercial health insurance plans created an incentive for many physicians to prescribe *abaloparatide* as the preferred therapy over its anabolic and antiresorptive competitors.

In terms of revenue performance and projected future growth, Radius's pricing strategy has also been a success. *abaloparatide* surpassed its revenue guidance of $95

to $98 million domestic net sales in 2018, and updated its 2019 revenue guidance to $165 to $170 million domestic net sales through October 2019. The pricing strategy of *abaloparatide* has proved to be financially sustainable to the drug manufacturer.

9.4.2 Challenges

However, the pricing strategy for *abaloparatide* also has encountered some challenges, most notably in expanding its Medicare Part D coverage. Twenty months after launch, *abaloparatide* has only achieved 67% coverage among Medicare Part D beneficiaries, compared to 94% coverage for *teriparatide*.

The launch of *abaloparatide* took place during the formulary review cycle of the Center of Medicare and Medicaid Services (CMS). This accident of timing was responsible for a delayed addition of *abaloparatide* to Part D formularies. This delay partly contributes to its low Part D coverage. The current structure of the Medicare Part D program creates a perverse incentive for Part D health plans to favor drugs with higher list prices in exchange for higher rebates. During their initial formulary negotiations with Radius, several health plans expressed concern over the financial disincentive caused by *abaloparatide*'s lower list price. Others, however, preferred *abaloparatide*, since it reduced the overall cost to the healthcare system.

Another challenge to Radius's pricing strategy is the uncomfortable fact that a lower list price does not necessarily lead to lower out-of-pocket costs for all patients. The out-of-pocket cost is set by multiple factors other than list price, including formulary tiers, copay and coinsurance payments, and insurance premiums. Even within the standard Medicare Part D plan, a patient may incur different levels of out-of-pocket costs at different phases of its coverage. For example, in the final catastrophic phase of Medicare Part D, *abaloparatide* has a 59% lower out-of-pocket cost than its main competitor. However, the out-of-pocket cost during the coverage gap before the catastrophic phase (the notorious "donut hole") is a heavy financial burden for many patients. Fifty percent of these patients will discontinue their treatments after seven months of their prescribed therapy, out of a recommended treatment period of eighteen to twenty-four months.

Radius also faces new competition in the anabolic agent market. In addition to biosimilar versions of *teriparatide* potentially launching in the second half of 2019, Amgen's novel anabolic agent romosozumab received FDA approval on April 9, 2019. While the presence of more agents will likely increase the patient volume of the anabolic agent market, it also inevitably puts pressure on the market share of *abaloparatide*.

Finally, as a small biotech startup with a relatively short history of commercial experience, Radius faces pressure to build brand loyalty among healthcare practitioners and to establish itself as a trusted partner with health insurance payers. Its single-drug portfolio limits its pricing flexibility, since it must generate revenue to support future R&D and create returns to its investors.

9.4.3 Future Evolution

The list price of *abaloparatide* increased by 5.9% on January 1, 2019. As of the time of submission of this article, biosimilar versions of *teriparatide* have not entered the U.S. market. The extent to which increasing competition will affect the market performance of *abaloparatide* remains to be seen. Radius Health, however, still remains committed to its socially responsible pricing strategy.

9.5 Conclusion

Radius Health launched *abaloparatide* at a list price 45% lower than its main competitor. This disruptive pricing strategy was based on several key factors, including the increasingly competitive landscape for women's osteoporosis treatment, the decline in the market volume for anabolic therapies, and a commitment to responsible drug pricing. Twenty months after its approval, *abaloparatide* has achieved nearly full coverage of the commercial and Medicaid segments of the U.S. health plan market, and over two-thirds of the Medicare Part D segment. It has captured a considerable share of the anabolic agent market, grown its patient volume, and caused Radius to exceed its revenue guidance to the benefit of its future sales and the R&D budget of

the company. Its pricing strategy has so far proven to be financially sustainable.

However, several challenges to this pricing strategy remain. A lower list price may not necessarily lead to lower out-of-pocket costs for all patients due to the structure of their health insurance, such as the "donut hole" in coverage for Medicare Part D. As a result, many patients prematurely discontinue their treatments. There is also a financial disincentive to certain health plans to place a drug with lower list price onto their formularies, since the rebate the plans will receive is lower than those of competing drugs.

There is growing public concern over incentives to health plans to favor drugs with higher list prices and higher rebates. Recent healthcare reforms are intended to implement changes that create incentives for health plans to adopt drugs that are priced responsibly. For example, the Department of Health and Human Services recently proposed to replace the rebate system with upfront discounts in drug list price, although the details of the proposed implementation remain to be seen.

It should also be noted that pricing is not the ultimate factor that determines the success of a novel drug or therapy. A pharmaceutical company seldom achieves market leadership by underpricing its competitors, but rather by using its pricing strategy to facilitate translational biomedical research and to create innovative products with improved therapeutic outcomes and drug delivery technologies. Nevertheless, socially responsible drug pricing creates numerous positive spillovers, reducing the overall cost to the healthcare system and benefiting a large group of patients under standard health insurance plans, while establishing brand loyalty among healthcare stakeholders that is particularly important to young commercial biotech companies like Radius Health.

This case study of *abaloparatide* illustrates both the success and the challenges of a fair and responsible drug pricing strategy, and potential applications to many other therapeutic areas. With continued reform and medical innovation in the healthcare industry, this method of pricing may be a highly effective way to simultaneously maximize the revenue of a pharmaceutical company and benefits to the patients.

Chapter 10

Review of Ethical Considerations of Human Challenge Trials

The COVID-19 pandemic has triggered an intense debate on the ethics of using controlled human challenge trials (HCTs), in which participants are inoculated with a pathogen, to accelerate the clinical development of vaccines and antiviral therapeutics. The benefits of HCTs must be weighed against the risks of deliberate infection of healthy individuals. In the past, HCTs have caused many tragedies due to negligence, a lack of informed consent, and the enrollment of subjects by manipulation and coercion. To inform the ethical considerations of ongoing and future HCTs, we provide a systematic review of the history of HCTs and examine the controversial issues in the ethical debate surrounding COVID-19 HCTs. We argue that the advent of mRNA vaccine technology and the quantitative modeling of infectious diseases will expedite the ethical assessment of future HCTs. Since delayed initiation significantly reduces the benefits of HCTs, it is critical for stakeholders to proactively establish standardized ethical criteria before the next pandemic so that future HCTs may be initiated with minimal delay if deemed ethical. [1]

[1]Joint work with Amanda Hu, Kien Wei Siah, Chi Heem Wong, and Andrew W. Lo. Valuable feedbacks from Jayna Cummings are gratefully acknowledged.

10.1 Introduction

Human challenge trials (HCTs), also known as controlled human infection model studies, have played an important role in the study of infectious diseases and vaccine development. By inoculating a small number of participants in a controlled experimental setting, HCTs enable a much more precise and systematic study of disease pathology. Compared to standard clinical trials, they may greatly accelerate the clinical development of vaccines and other anti-infective therapeutics. The scientific results of HCTs are also more informative regarding the testing the safety and efficacy of vaccines than preclinical data of animal studies [174]. These unique advantages of HCTs are especially appealing during a global pandemic, when vaccines and treatments are direly needed to prevent enormous casualties and socioeconomic losses. Since the start of the COVID-19 pandemic, some bioethicists and biomedical researchers have strongly advocated for HCTs to accelerate vaccine and therapeutic development [175]–[177].

However, the advantages of HCTs must be weighed against serious ethical concerns about the deliberate inoculation of healthy subjects with a potentially debilitating or lethal pathogen. These ethical concerns stem from horrifying incidents in the history of HCT, where researchers performed inoculations haphazardly without understanding the virulence of the disease, and disadvantaged individuals were manipulated, deceived, or even coerced to be infected with deadly pathogens. Though these tragedies led to the establishment of statutes to protect the participant's right of informed consent and to ensure that the experimental risk/benefit tradeoff is ethically acceptable, currently there are no widely accepted standardized guidelines which specify the conditions under which HCTs are deemed ethical. During a pandemic caused by a previously unknown pathogen, it will take time for regulators to produce ad hoc ethical guidelines for HCTs and convene expert panels to examine proposals for HCTs at length. Even though regulatory scrutiny is required, delay in initiating a HCT undermines its benefit to accelerate vaccine development in a rapidly evolving pandemic. Partly for this reason, HCTs have not yet been used to produce effective

vaccines or treatments for COVID-19. To help inform future regulatory decisions on HCT, we provide a systematic review of the history of HCTs, explicate their ethical controversies, and examine the ongoing ethical debate over the use of HCTs for COVID-19. We contribute two new perspectives from the current pandemic on how breakthroughs in vaccine technology and quantitative modeling may facilitate the ethical assessment of HCTs. We conclude with a call to action for stakeholders to proactively establish standardized and practical ethical criteria for the use of HCTs to prevent the next pandemic.

10.2 Early History of HCTs

The history of HCTs consists of both groundbreaking scientific advances and atrocities against humanity. The first HCT in modern history occurred in the eighteenth century, though it is probable that the intentional inoculation of healthy subjects in China, India, and Africa occurred much earlier [178]. We review the history that is most relevant to shaping the current ethical debate on HCTs for COVID-19. For a comprehensive review of HCT history, see [179].

Early HCTs would be considered unethical by today's standards, yet they represent a significant advance in the study of infectious disease. Edward Jenner's invention of smallpox vaccination in 1796 is the first recorded HCT in modern history [180]. Having learned that dairymaids were protected from smallpox after they were naturally infected by cowpox, Jenner tested whether cowpox provided immunity to smallpox by inoculating an eight-year-old boy with cowpox, and nine days later, with smallpox. The boy showed no signs of smallpox, leading to the first modern vaccination. In his publication recording the study, Jenner coined the word "vaccination" from *vaccinia*, the Latin word for cowpox.

In 1885, Daniel Alcides Carrión, a medical student in Peru, conducted one of the first HCTs via self-experimentation [181]. His goal was to prove the link between *verruga peruana*, an endemic yet non-infectious disease, and Oroya fever, a severe infectious disease affecting hundreds of railroad workers in Peru at the time. After

his supervisors refused his request for self-experimentation, Carrión inoculated himself with materials from the skin lesion of a patient affected by *verruga peruana*, developed severe symptoms of Oroya fever, and died five weeks after inoculation. His self-experiment proved that Oroya fever is the acute form of the chronic *verruga peruana*. Although Carrión is honored as a medical martyr and national hero of Peru, the ethical justification of self-experimentation by physicians remains highly controversial.

The Cuba yellow fever trial in 1900, led by Major Walter Reed of the U.S. Army, is another early HCT that caused significant ethical controversy [182]. Reed's research team inoculated soldiers and physicians with mosquitoes to test whether mosquitos transmit the disease. The inoculation was performed with neither Reed's supervision, nor any clear protocols for participant selection. Initially, no infections were observed, and the potential risks of HCT were increasingly overlooked. Suddenly, two young physicians who had self-inoculated died of the disease. After that, Reed returned to Cuba to supervise the research and design new protocols for the HCT. Meanwhile, additional ethical issues were raised as native Cubans and Spanish immigrants were recruited as participants in an unjust manner, offering substantial compensation, which suggests possible manipulation and exploitation. In total, Reed's yellow fever study resulted in four deaths.

The first successful well-documented HCT was conducted for influenza in 1937 in the Soviet Union [183]. About 20% of its 72 participants exhibited mild symptoms after inoculation, and no death or severe infection was reported. This study established the safety of HCTs for influenza, which led to subsequent HCTs to test influenza vaccines and treatments. Despite its scientific significance, it is difficult to assess the ethical justification of this trial, since its protocols for participant selection, compensation, and informed consent have not been made available.

The ethical issues around haphazard study protocol, researcher negligence, and the recruitment of manipulated or coerced participants from disadvantaged and discriminated populations were exacerbated in the worst years of the twentieth century. During World War II, Nazi Germany and Imperial Japan conducted human challenge studies on prisoners of war and civilians in occupied areas, causing enormous casual-

ties [184]. These atrocities led to the creation of the Nuremberg Code in 1947, which codifies ten fundamental principles of ethical research involving human subjects, including voluntary consent, the avoidance of unnecessary suffering and injury, and the principle that the risk to the participant "should never exceed that determined by the humanitarian importance of the problem to be solved by the experiment." These became the basis for the assessment of ethical justifications of HCTs thereafter.

10.3 HCTs since Nuremberg

In the years after 1947, governments and international organizations gradually enacted statutes to enforce the principles outlined in the Nuremberg Code. Some HCTs during this time were still conducted with unacceptable ethical standards, especially in underdeveloped regions. One notorious case is the Guatemala syphilis experiment, funded by the U.S. National Institutes of Health (NIH), in which 1,308 individuals, including prisoners and mentally incapacitated patients, were inoculated with sexually transmitted diseases without informed consent. Eighty-three of these individuals died because of the study [185].

In the decades following, HCT gradually became ethically acceptable, thanks to advances in biomedicine which led to effective treatments as rescue therapies, and regulations that protected the rights of participants and mandated institutional review boards to ensure that the HCT protocol would meet ethical guidelines before any subjects have been enrolled. Ethical HCTs have been conducted for common infectious diseases such as influenza, malaria, typhoid, cholera, and dengue fever, among others. In the twenty-first century, there have been over 50 successful phase 1 or phase 2 HCTs for these five diseases (Figure G-1). HCTs have helped identify many effective vaccines and treatments, including the first antiviral drug and the first live attenuated vaccine for influenza, the Vaxchora vaccine for cholera, the Typbar-TCV vaccine for typhoid, and the RTS,S vaccine for malaria [186]–[189]. A study in 2018 estimates that there have been 500 recorded HCTs since 1970, targeting more than 20 diseases [190].

The ethical justifications for HCTs in general fall into two categories. In the first case, HCTs may be justified if there exist rescue therapies to cure the disease, and the risks of severe illness and mortality are extremely low. A notable example is influenza. A recent review argues that the HCT is the essential clinical study design to advance vaccine and therapeutic development for influenza, and can be conducted safely [191]. While adverse events are rare in influenza HCTs, one incident may suffice to severely undermine public opinion of its ethics. In 2000, a 21-year-old participant experienced a cardiac event after an HCT [192]. Although researchers could not prove that the HCT caused the cardiac event, influenza HCTs in the U.S. were suspended until 2014, when researchers at NIH ran a successful dose-finding HCT in response to the 2009 H1N1 pandemic [193]. For highly lethal diseases with no rescue therapy, such as Ebola, HCT cannot be ethically justified, regardless of its potential social or scientific value, or whether participants provide informed consent [194].

If there is substantial uncertainty in the pathogenesis, transmission, and virulence of the disease but no rescue therapy is available, the ethical justification of HCTs becomes much more complex. Two necessary conditions are (1) the perceived benefits to the society must outweigh the risks to the participants, and (2) an HCT is the only option to achieve these benefits. During the Zika virus outbreak in 2015-16, the U.S. National Institute of Allergy and Infectious Diseases (NIAID) issued a report that recommended against a proposal for HCTs to study the pathogenesis and immune responses of Zika [195], [196]. The report argued that there was a lack of strong evidence for a social benefit in anticipated results of an HCT to expedite vaccine development for the Zika virus. In addition, the endpoints of an HCT could be achieved via a standard trial. Furthermore, an HCT in this case would generate significant risks to third parties outside the trial, since the Zika virus can be transmitted sexually. Nonetheless, the report recommended reconsidering HCTs under future conditions—without, however, specifying any quantitative risk or benefit metrics of the HCT to be met in order to receive ethical approval. This lack of widely accepted standardized criteria to measure the tradeoffs of HCTs caused many controversies in the ethical debate of applying HCTs in COVID-19 research.

10.4 The Ethics of COVID-19 HCTs

Given the significant numbers of casualties and the socioeconomic loss caused by COVID-19, some bioethicists strongly advocated for HCTs to expedite the development of vaccines since the early months of the pandemic [175]–[177]. On May 3, 2020, a WHO Working Group report stated that HCTs for COVID-19 research are ethically acceptable under eight criteria [83]. These criteria span four areas: scientific and ethical assessment, consultation and coordination, participant and site selection, and review by experts and informed consent by participants. In the ensuing ethics debate, both proponents and opponents of HCT accepted these criteria as necessary conditions for an ethical HCT, but the opponents strongly challenged whether these criteria (and their underlying assumptions) were sufficient to justify the ethics of HCT. We summarize the six key areas of ethical debate in Appendix G.1 and review the four most controversial issues below.

The most important ethical justification for HCTs is that they accelerate vaccine development by testing a much smaller group of participants than standard vaccine trials. One study acknowledges that if standard trials progress sufficiently rapidly, HCTs will not be necessary [176]. However, opponents argue that HCTs require more careful ethical review and elaborate technical preparation, such as producing virus strains with attenuated virulence in Good Manufacturing Practice (GMP) facilities and engaging the local community at HCT sites to establish the study protocol. The process is further prolonged by the need to conduct a dose-escalation study before testing the main endpoints, and a follow-up trial to test the safety and efficacy on a larger population [197]. Some experts estimate that it will take 1 to 2 years to set up a COVID-19 HCT, making it unlikely to accelerate vaccine development [198]. Holm further argued that the social value of accelerated vaccine development is overestimated due to the lack of equitable access to vaccines, especially in low- and middle-income regions, a prediction verified by reality [199].

Another controversial issue is the scientific value of HCTs. The controlled inoculation in an HCT enables a much more precise and systematic study of the patho-

genesis, transmission, and immune response to COVID-19, which may inform public health policy on nonpharmaceutical interventions (e.g., social distancing mandates). However, its scientific value is inherently limited by sample selection bias, since only young, healthy adults may be enrolled in an HCT to minimize the risks of infection. It is difficult to generalize these study results to high-risk populations (e.g., elderly or immunocompromised populations, or those with comorbidities) who will benefit the most from vaccines [175], [198], [200], [201]. A follow-up standard vaccine trial to test the safety and efficacy on these high-risk populations is required to receive FDA approval [197].

In addition, several studies argue that the ethical basis of informed consent by the participants is undermined by the highly uncertain and continuously evolving risks of COVID-19 and long-term post-COVID conditions [197], [201], [202]. Additionally, the participants may be misled by "preventive misconception," the false belief that participation in an HCT alone will provide some level of protection against COVID-19. The opponents argue that HCT participants cannot genuinely grant informed consent in this context, since they will overestimate the benefits and underestimate the serious risks of infection and post-infection syndromes.

The potential fairness of the HCT site and participant selection process has also been challenged. Proponents of HCTs argue that participants should be restricted to those who are at substantial risk of natural exposure to COVID-19 [177]. Such individuals face smaller marginal risks from participating in an HCT, and may benefit by the immunity acquired in an HCT. Adequate financial compensation further reduces their "net risk" of participation in an HCT. However, socioeconomically disadvantaged populations are more heavily affected by COVID-19, and thus are more likely to be attracted by the substantial compensation of an HCT, raising concerns of exploitation and manipulation [199], [201], [202]. These populations bear a disproportional burden of disease incidence, while an HCT would divert the scarce medical resources needed in their communities [197].

10.5 Lessons from COVID-19

We raise several new perspectives regarding the ethical consideration of HCTs learned from the current pandemic, and a call for action. First, breakthroughs in vaccine development for COVID-19, such as mRNA technology, have changed the prospective conditions for the use of HCTs. The first dose of mRNA-1273 vaccine was administered in its phase 1 clinical trial on March 16, 2020, only 64 days after the DNA sequence of COVID-19 was made available. The U.S. FDA authorized the vaccine for emergency use nine months after that, while the first HCT had still not been initiated by that time. Nonetheless, the advent of mRNA vaccines does not doom the use of HCTs, since the demonstrated safety and efficacy of mRNA vaccines may also reduce the risks of future HCTS and increase their probability of success. It is conceivable that we may get the best of both worlds through combining this novel vaccine technology with HCTs to tackle new variants of COVID-19, test novel delivery mechanisms, and prevent future pandemics.

In addition, the pandemic has triggered an explosion in the quantitative modeling of infectious diseases, both at the patient level, in diagnosis and prognosis, and at the population level, in epidemiological forecasting. These advances will inform the ethical assessment of HCTs by supplementing qualitative, principle-driven arguments with precise, data-driven recommendations. Although they should not replace human decision-making, these quantitative models may provide rational, rigorous, and transparent frameworks to evaluate the risks and benefits of HCTs (and other clinical trial designs) under different scenarios, and determine whether an HCT will achieve the optimal risk/benefit tradeoff among all clinical trial designs under consideration.

Berry *et al.* perform the first quantitative risk/benefit analysis comparing an HCT design with three alternative trial designs for COVID-19 vaccines: standard randomized clinical trial (RCT), optimized vaccine efficacy RCT (ORCT), and adaptive RCT (ARCT) [34]. The authors simulate vaccine development under COVID-19 (starting on August 1, 2020 in the U.S.) using an epidemiological model which accounts for social distancing and vaccination. The model makes specific assumptions about epi-

demic, vaccine, and public health policies (e.g., infection and mortality rate, vaccine efficacy, vaccination rate, etc.), which can be adjusted to simulate different scenarios. The risk/benefit tradeoff of each trial design is the expected number of infections and deaths prevented by a vaccine developed with this trial minus the number of infections and deaths of participants in the trial.

This analysis shows that HCTs have the optimal risk/benefit tradeoff if one can be set up in 30 days. This result holds for different vaccine efficacies (from 30% to 90%) and vaccination scenarios. However, if the HCT requires more than 60 days to set up, it causes more infections and deaths than an ARCT (which also holds in different scenarios), making an HCT unethical in this case. While the analysis's conditional recommendation of an HCT agrees with [198], the quantitative model of [34] draws a practical boundary between when it is ethical to use HCTs (a 30-day set-up) and when it is unethical (a 60-day set-up). It can be easily adapted to evaluate the risk/benefit tradeoffs during future epidemic outbreaks at different localities.

Finally, since the benefit of an HCT to accelerate vaccine development is reduced if the HCT cannot be initiated in a timely manner, it is critical to minimize the decision time to assess the ethical justifications during a pandemic. An effective solution is to establish standardized ethical criteria for HCTs well before the next pandemic to avoid the delay caused by drafting ad hoc ethical guidelines. A successful example of such a standardized ethical criterion is the legislation which allows physicians to withdraw life support from patients who have signed a do-not-resuscitate order or have designated a surrogate to make such requests, reducing the prolonged suffering of the patient, the financial loss of the family, and the ethical burden of the physician [203]. During a rapidly evolving pandemic, the indecision of public health regulators can cause as much harm as a bad decision, and an ethical HCT may become unethical due to delay. Therefore, stakeholders must be proactive to establish ethical criteria and build the technical infrastructure so that a future HCT can be initiated with minimal delay if deemed ethical.

By the time of writing this manuscript, there are two ongoing HCTs for COVID-19, both in the United Kingdom (UK) [204], [205]. One trial (NCT04865237), funded by

the UK government and initiated in March 2021, is evaluating the efficacy of vaccine and treatment candidates. The other trial (NCT04864548), funded by the Wellcome Trust and initiated in May 2021, is studying the immune response to COVID-19 after reinfection. So far, no incidents of severe illness or death have been reported. We hope our review will help the readers make an informed judgment on the ethics of the ongoing and future HCTs for COVID-19 and other diseases.

10.6 Conclusion

The HCT is an efficient clinical trial design which has led to both life-saving therapeutics and, when conducted unethically, horrible tragedies. The ethical justification for HCTs hinges on whether its perceived benefits to society outweigh the potential risks to its participants, a difficult quantity to gauge when there are no standardized ethical criteria. The COVID-19 pandemic has triggered ethical debates regarding the scientific value, social benefit, feasibility, and fairness of participant selection of HCTs. Given mRNA vaccines and quantitative models to estimate the risks and benefits of an HCT, future HCTs may have higher probabilities of success and more robust ethical justifications supported by rigorous evaluations of the risk/benefit tradeoff. To maximize the benefit of an HCT in accelerating vaccine development, stakeholders must proactively establish standardized ethical criteria before the next pandemic, and minimize unnecessary delay to initiate the HCT if it is deemed ethical.

Part VI

Conclusion

Chapter 11

Summary of Findings

In this thesis, we propose various financial and analytical strategies to facilitate translational biomedical research and novel drug development. We conclude with a summary of the key findings and insights from each part of the thesis.

11.1 Summary of Part II

The pair of simulation studies of the biomedical megafund model forms an interesting contrast and illustrates the advantages and limitations of the portfolio approach of financing drug development. The megafund for glioblastoma (GBM) therapeutics generates an attractive annualized return of 14.9% on average despite the low probability of success (PoS) of drug development outcomes for GBM. The megafund for mRNA vaccines, on the other hand, does not generate financial value under a wide range of assumed parameter values (unless the price per vaccine dose is $78.00 or above) despite the much higher PoS of mRNA vaccines.

Multiple factors contribute to the difference in financial performance. First, the GBM megafund is diversified across different therapeutic mechanisms and clinical phases, while all assets in the vaccine megafund have the same therapeutic mechanism and initial clinical phase, leading to larger financial risks. Second, the adaptive clinical trial platform GBM AGILE significantly reduces the costs and duration of late-stage clinical trials for GBM drug candidates, while the phase 3 trials of the

mRNA vaccines cost $150 million each and constitute 59% of the total investments. Most importantly, the "multiple shots on goal" approach of parallel drug development effectively lowers the supply side risk by increasing the PoS of having at least one successful drug approval in the portfolio. However, it does not mitigate the demand side risk for drug sales and revenue, which is especially significant for vaccines due to the stochastic nature of epidemic outbreaks. As a result of the demand side risk, the revenue generated by the FDA-approved vaccines is insufficient to recover the costs of clinical trials, while the revenue generated by an approved GBM therapy is much more stable. Our analysis shows that the portfolio approach alone does not guarantee positive financial value for the investors if the demand side risk remains large. In the case of the vaccine megafund, greater public-private partnership and and more cost-efficient clinical trial designs (discussed in Part III) are needed to ensure that vaccine development is a financially sustainable business model for the pharmaceutical companies.

11.2 Summary of Part III

The novel clinical trial design based on Bayesian decision analysis (BDA) provides a rational, quantitative, transparent, and adaptable framework to incorporate the patients' value and preference in the FDA's regulatory decision under uncertainty and complex tradeoffs. The patients' willingness to accept a higher risk of adverse effects in exchange for expedited approvals of potentially effective therapies is reflected in the Bayesian optimal Type I error rate, often higher than the traditional 2.5% or 5% threshold required by the FDA. For vaccine development under a rapidly evolving epidemic outbreak, the BDA model recommends a smaller patient size and a higher Type I error rate when the disease is more infectious and fatal. For the controversial new drug application of AMX0035 treating the fatal disease amyotrophic lateral sclerosis (ALS), the BDA model recommends that the FDA should approve the drug candidate based on its phase 2 trial results since the potential therapeutic benefits outweigh the limited risks of adverse effects, which is consistent with the patient values expressed

by the ALS advocacy groups [94]. Although the BDA framework should not replace human judgement and does not capture the full complexity of FDA's regulatory decision, it serves as a useful reference which provides concrete recommendations for the FDA when balancing complex tradeoffs under large uncertainty.

11.3 Summary of Part IV

We apply statistical techniques and novel machine learning models to estimate three key parameters of the drug development process: the probability of success (PoS) of a drug approval, the duration of a clinical trial, and the correlations between clinical trial outcomes. Using the Debiasing Variational Auto-Encoder (DB-VAE) for PoS prediction reveals significant bias present in the machine learning models trained on the highly imbalanced dataset of historical drug approval outcomes. Debiasing increases the overall prediction performance and generates financial value for the drug developer. Predicting clinical trial duration reveals important factors which impact the trial duration, such as the type of trial sponsor and the therapeutic area. Estimating the correlation using biostatistical techinques reveals positive yet weak correlations (below 8%, p-value <0.001) between clinical trial outcomes in each therapeutic area, though the specification of correlation structure should be further improved to capture the complex biomedical factors which induce correlation while maintaining computational feasibility. Overall, our studies show that applying big data analytics to analyzie historical drug development data may provide many important and novel insights for managing the biomedical megafund and designing clinical trials.

11.4 Summary of Part V

The financial and analytical innovations discussed in the previous parts are based on abstract models of drug development which capture certain aspects of the drug development process but do not consider the practical challenges when these innovations are implemented in the real-world healthcare system. The two studies in Part V

illustrate the practical challenges of implementing a novel drug pricing strategy and a controversial human challenge trial (HCT) design. For the disruptive pricing strategy of *abaloparatide* with 45% lower list price than its main competitor, the challenges come from a perverse incentive of certain health plans which favor higher list prices in exchange for higher rebates from the drug manufacturer. For using the HCT to expedite vaccine development for COVID-19, the challenge largely comes from the absence of well established ethical criteria and robust quantitative models to gauge the risk/benefit tradeoff of conducting an HCT under different hypothetical scenarios. As a result, the first HCT for COVID-19 were initiated after the first vaccine candidates had already received emergency use authorization from the FDA in U.S. and Europe. The main take-away is that the innovative strategies must be accompanied by regulatory support in order to truly innovate the drug development process and benefit the patients. It is critical to closely collaborate with the regulatory agencies and other stakeholders to establish the conditions under which these novel solutions can and should be properly applied to improve various aspects of the drug development and healthcare system.

Part VII

Appendix

Appendix A

Supplements to Chapter 2

A.1 Methods

In this chapter, 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 [11], a pediatric oncology megafund [31], and an ovarian cancer portfolio [30] (see Figure S1). 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 [54]. As an inferentially seamless phase 2/3 platform trial [206], 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 [2].

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.

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 approaches, 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 acquired through years of domain-specific experience. In this paper, we identify scientifically promising approaches 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 portfolio (Table 2.1). 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 the profitability of approved drugs.

Probability of Success, Cost of Development, and Duration

We first compiled 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 [2], [30], [207]–[209]. Next, we asked each expert from the NBTS network to estimate the same set of parameters based on their experience. To reduce the impact of outliers, we focused on the median of the estimates provided by the panel. Finally, we took 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.

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 Results section, we perform a sensitivity analysis of our results with respect to this factor.

GBM AGILE

GBM AGILE is a global, two-stage platform trial [54] designed to facilitate the expedited approval of effective therapies for GBM, and reduce the cost in performing large-scale clinical studies [54]. It is operated by the Global Coalition for Adaptive Research (GCAR), a 501(c)(3) nonprofit organization [53]. A platform trial evaluates the effects of multiple therapies, each as an experimental arm, against a common control arm. The platform is maintained perpetually under a master protocol, and therapies enter or exit the platform based on a decision algorithm. 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 probability of receiving each therapy proportional to

the Bayesian probability of that therapy improving overall survival, which is updated monthly. Promising therapies identified in the first stage then seamlessly transition to the second stage (the "confirmation stage"), 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 into more promising arms is prioritized, better therapies therefore 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 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 of GBM AGILE 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 Figure S2). To simulate the uncertainty in the project selection process, we assume that each asset in the megafund portfolio that is eligible for GBM AGILE has a probability p_{inc} of being included in the platform. The assets not included in 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 therapeutic effects earlier will enter stage 2 with a smaller number of accrued patients ("early graduation")—further reducing the cost and duration of these trials. The 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 3 trials for assets exiting 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 Supplementary Materials 4).

We derive our cost and duration estimates assuming a steady state of one control

arm plus three experimental arms in GBM AGILE. We calibrate our estimates—patient accrual rate, cost per patient, probability of inclusion and graduation of each stage—with the input from both NBTS and GCAR network of experts. No literature estimates were available at the time of this study, since GBM AGILE is the first global, disease-specific platform trial for GBM [54]. We note that the cost and duration of each GBM AGILE trial are much lower than standard phase 2 and phase 3 trials combined, about 75–85% lower in terms of cost, and 20–30% shorter in terms of duration.

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 Figure S3) before projecting the resulting correlation matrix to its nearest positive-definite counterpart for use in our simulations [62]. See Supplementary Materials 4 for the details of implementation.

Profitability of a Successful 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, however, it is difficult to estimate the profitability of an approved brain cancer drug, as [30] 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 [30] may not be appropriate in this analysis.

Instead, we follow the approach used in [11] and [31]. 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"—relative to other treatments. We calibrated our assumptions through discussions with the NBTS network of experts, a review of the current standard of care, and market research reports [210]. In our base case, the NPV of approved drugs in our portfolio ranges between $530 million and $2.988 billion, with a median valuation of $1.272 billion. Figure S4 illustrates the investment timeline of a drug in our portfolio.

A.2 Cost and Duration of GBM AGILE

We derive cost and duration estimates for GBM AGILE using the assumptions of a steady state with 4 arms (1 control and 3 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 3

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:

$$d = \frac{n}{\frac{ves}{m}} + f \tag{A.1}$$

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:

$$c_e = pm \text{ and } c_c = \frac{p}{m}(1-e)n \tag{A.2}$$

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., \$0.084M). 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 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%.

In the Results section of the main text, we perform a sensitivity analysis of our results with respect to patient accrual per month and number of experimental arms in steady state.

A.3 Correlation

We first average the correlation estimates made by the experts. Next, we symmetrize the resulting correlation matrix by performing the following operation:

$$R := \frac{1}{2}\big(X + X^T\big) \tag{A.3}$$

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 [62].

To generate correlated trial outcomes, we first draw a vector of random multivariate standard normal variables $\epsilon_j := [\epsilon_{1j}, \epsilon_{2j}, ..., \epsilon_{nj}]^T$, 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 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, we can model trial outcomes as Bernoulli variables $B_{ij} = \mathrm{I}\{z_{ij} > \alpha_j\}$, 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, p_j is the probability of success for phase j.

A.4 Computing the Annualized Rate of Return

The GBM megafund portfolio has a complex cash flow structure. Large amounts of investment are due at the beginning of each clinical trial phase for each asset and, when an asset is included in GBM AGILE, the costs of new patient accrual are due each quarter. In addition, since the assets are diversified across different initial stages of development, some assets may have received regulatory approval and begin to generate revenue while others remain in the pipeline. To the best of our knowledge, there is no standard approach to compute the annualized rate of return R_a with multiple-period revenue and investment.

We adopt the approach of [30] to compute R_a in Tables 2.2, 2.3, 2.4. To provide readers with transparency and additional insights into our calculations of the financial risk and reward of the GBM megafund, we discuss three alternative approaches to define R_a, and provide the corresponding simulation results in this section.

Let CF_t denote the cash flow of the GBM venture fund at time $t \in T = \{t_0, ..., t_n\}$. The total revenue X and investment cost C are given by

$$X = \sum_{t \in T} CF_t \mathrm{I}\{CF_t > 0\}, \quad C = \sum_{t \in T} |CF_t| \mathrm{I}\{CF_t < 0\} \tag{A.4}$$

Following [30], we define the annualized rate of return R_a as the annualized arithmetic mean (AM) of the cumulative return

$$R_a^{AM} := \frac{1}{T}\Big(\frac{X}{C} - 1\Big) \tag{A.5}$$

where T denotes the horizon of the megafund investment (i.e. the time of the last nonzero cash flow). Alternatively, one can define R_a as the annualized geometric mean (GM) of the cumulative return

$$R_a^{GM} := \Big(\frac{X}{C}\Big)^{1/T} - 1 \tag{A.6}$$

The two definitions of R_a described above are intuitive, and simple to convert to cumulative return. However, they do not apply a time discount to the cash flows, and a cash flow may have negative NPV but positive R_a. To account for the time value of money with annual discount rate r, one can define the time-discounted versions of revenue and investment costs

$$X_d = \sum_{t\in T} \frac{CF_t}{(1+r)^t}\mathrm{I}\{CF_t > 0\}, \quad C_d = \sum_{t\in T} \frac{|CF_t|}{(1+r)^t}\mathrm{I}\{CF_t < 0\} \tag{A.7}$$

Under this definition, the NPV of the megafund portfolio is given by

$$NPV = X_d - C_d \tag{A.8}$$

One can similarly define the time-discounted arithmetic mean (AM-DC) and geometric mean (GM-DC) versions of the annualized rate of return

$$\begin{aligned} R_a^{AM-DC} &= \frac{1}{T}\Big(\frac{X_d}{C_d} - 1\Big) = \frac{1}{T}\frac{NPV}{C_d} \\ R_a^{GM-DC} &= \Big(\frac{X_d}{C_d}\Big)^{1/T} - 1 = \Big(\frac{NPV}{C_d} + 1\Big)^{1/T} - 1 \end{aligned} \tag{A.9}$$

For any cash flow, the NPV will always have the same sign as R_a^{AM-DC} and R_a^{GM-DC}. However, the expected NPV may still have a different sign from the ex-

pected R_a^{AM-DC} and R_a^{GM-DC} since the NPV, C_d and T are correlated random variables.

A.5 Value of Stage 1 Results of GBM AGILE

Stage 1 of GBM AGILE is intended to identify any clear therapeutic effects of therapies on target patient subtypes. Even if the therapy does not meet the criteria to graduate to stage 2 of GBM AGILE, the results from stage 1 may still provide useful guidance for biopharma companies to improve their drug trial design and continue its development via phase 2 or 3 trials in the standard 505(b)(1) pathway. To model this scenario, we assign a probability $p_{S1-P2/3}$ that the megafund continues development for each asset that shows promising stage 1 results but does not meet the criteria to enter stage 2. The baseline model assumes a $p_{S1-P2/3} = 20\%$, and a larger value of $p_{S1-P2/3}$ can be interpreted as an observation of more valuable results in stage 1. Varying this probability from 0 to 40%, we find that the expected annualized return rises by 3.9 percentage points, the probabilities of loss and wipeout drops by 7.1 and 7.3 percentage points, respectively, and the expected NPV grows by a factor of 4.

We also assign a probability $p_{P2/3-P3}$ that the megafund conducts a subsequent phase 3 trial for the asset instead of repeating phase 2. The baseline model assumes a $p_{P2/3-P3} = 50\%$. Varying this value from 0 to 100%, we observe similar improvements on the performance of the megafund as before. We conclude that GBM AGILE provides a financially efficient means to garner valuable therapeutic information on an asset that may guide future developments undertaken by the megafund outside the GBM AGILE platform.

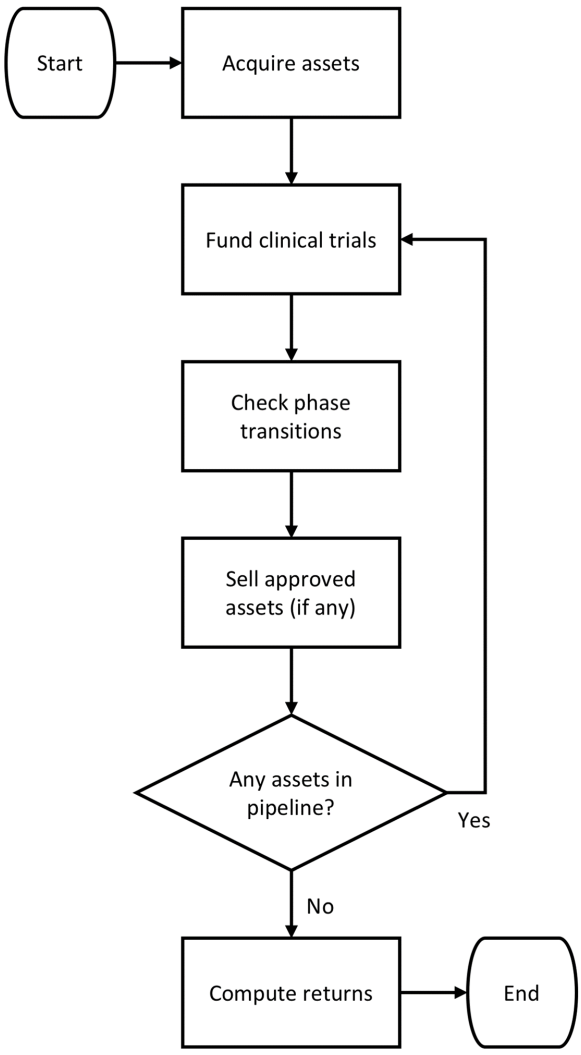


Figure A-1: Simulation framework for the brain tumor megafund.

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.

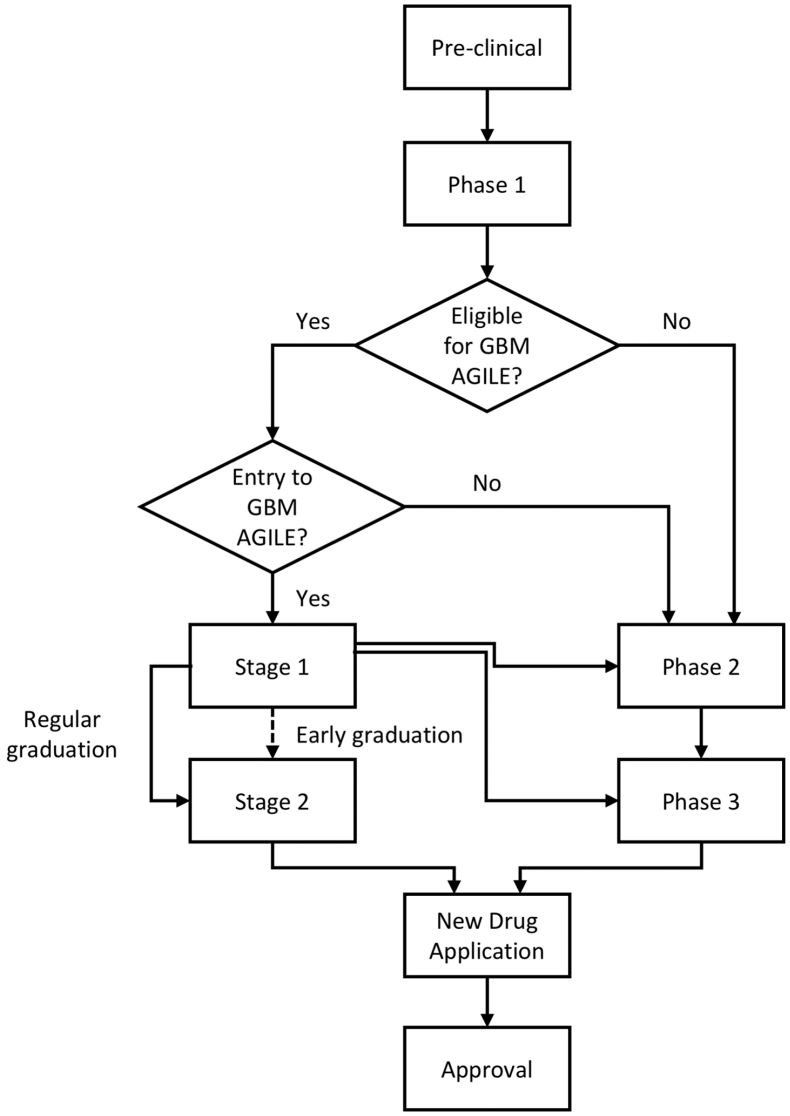


Figure A-2: Possible development paths for assets in the NBTS portfolio.

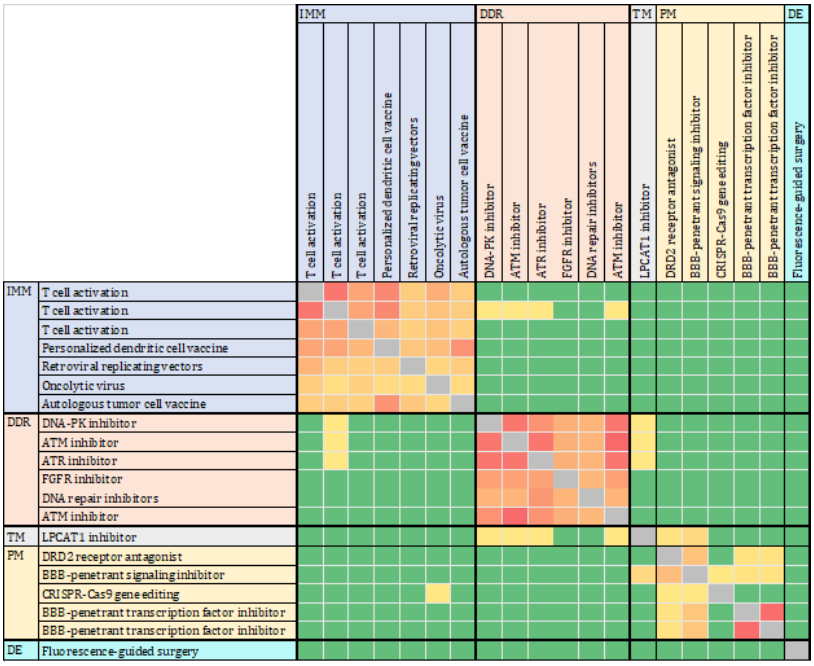


Figure A-3: Correlation matrix of brain cancer projects (average of estimates from all the NBTS network of GBM experts).

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.

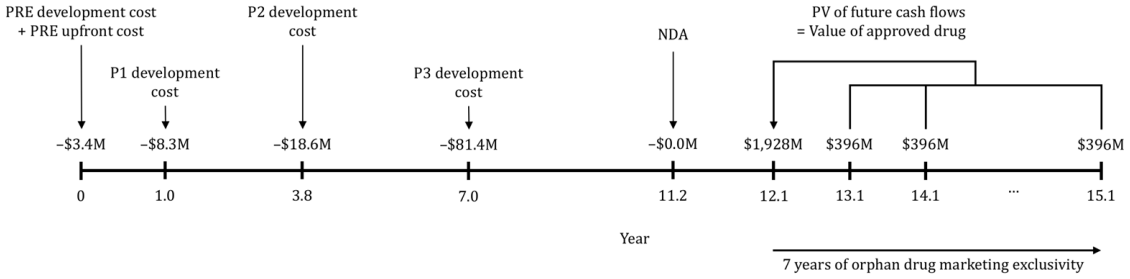


Figure A-4: Investment timeline of a brain cancer drug targeted at recurrent glioblastoma patients.

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 [211]. 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 patient15, 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 2x, i.e., $132,000, because it has been identified to be a transformative treatment by the experts. Assuming a 10% cost of capital, 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.

Table A.1: Literature estimates of model parameters for standard clinical trials.

Parameter	PRE to P1	P1 to P2	P2 to P3	P3 to NDA	NDA to Approval	PRE to Approval
Probability of success (%)	69.0	81.4	30.5	26.5	100.0	4.5
Duration (months)	12.0	42.1	40.6	48.5	9.6	152.8
Development cost ($M)	3.5	10.1	20.7	92.8	0.0	127.1
Discount factor (%)	23.0	20.0	17.2	12.5	10.0	NA

Table A.2: Model parameters estimated by the NBTS network of experts.

Parameter	PRE to P1	P1 to P2	P2 to P3	P3 to NDA	NDA to Approval	PRE to Approval
Probability of success (%)	60.0	60.0	40.0	45.0	100.0	6.5
Duration (months)	NA	24.0	38.0	51.5	12.0	NA
Development cost ($M)	1.0	6.5	16.5	70.0	NA	NA

Table A.3: Probability of success, costs of development, and duration at each phase of development for standard clinical trials.

Parameter	PRE to P1	P1 to P2	P2 to P3	P3 to NDA	NDA to Approval	PRE to Approval
Baseline PoS (%)	64.5	70.7	35.3	35.8	100.0	5.7
NBTS portfolio PoS (%)	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 ($M)	1.1	8.3	18.6	81.4	0.0	111.
Discount factor (%)	23.0	20.0	17.2	12.5	10.0	NA

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: PRE: preclinical; P1: phase 1; P2: phase 2; P3: phase 3; NDA: new drug application; NBTS: National Brain Tumor Society.

Table A.4: Probability of transition, costs of development, and duration at each stage of development for GBM AGILE.

Parameter	P1 to S1	S1 to S2 (early)	S1 to S2 (regular)	S1 to P2	S1 to P3	S1 to Futility	S2 to NDA
Probability of transition (%)	33.0	15.0	5.0	10.0	10.0	60.0	50.0
Duration (months)	7.0	21.0	15.0	21.0	21.0	21.0	42.0
Development cost ($M)	NA	15.2	10.7	15.2	15.2	15.2	7.5

Abbreviations: P1: phase 1; S1: stage 1; S2: stage 2; P2: phase 2; P3: phase 3; NDA: new drug application.

A.6 Estimating the net present value of approved drug candidates

We estimate the sales revenue for each drug candidate in the venture fund portfolio. The annual cost of Temodar, $66,000 per patient [15], adjusted to 2019 dollars using the Biomedical Research and Development Price Index, is used as a benchmark. We note that 14 drugs in the portfolio are priced at a premium of 2x, i.e., $132,000, because they are identified to be a transformative treatment by the NBTS network of experts. The drug candidates appear in the same order as in Table 2.1 of the main text. Net Present Value (NPV) is computed using a cost of capital 10%.

Table A.5: Estimating the net present value of successful drug candidates.

Market size (per year)	Market penetration (%)	Price per patient ($M)	Revenue per year ($M)	Marketing exclusivity (years)	Pediatric extension (years)	Priority review voucher ($M)	NPV ($M)
30,000	10.0	0.132	396.0	7.0			1,928
30,000	10.0	0.132	396.0	3.0			985
30,000	10.0	0.132	396.0	7.0			1,928
16,500	10.0	0.132	217.8	7.0	0.5	100.0	1,214
16,500	10.0	0.132	217.8	7.0	0.5	100.0	1,214
30,000	10.0	0.132	396.0	7.0			1,928
16,500	10.0	0.132	217.8	7.0			1,060
9,900	20.0	0.132	261.4	7.0			1,272
9,900	20.0	0.132	261.4	7.0			1,272
16,500	10.0	0.132	217.8	7.0			1,060
30,000	10.0	0.132	396.0	7.0			1,928
9,900	20.0	0.066	130.7	7.0			636
3,500	20.0	0.132	92.4	7.0	0.5	100.0	572
46,500	10.0	0.066	306.9	7.0			1,494
9,000	20.0	0.132	237.6	7.0	0.5	100.0	1,315
16,500	10.0	0.066	108.9	7.0			530
46,500	10.0	0.132	613.8	7.0			2,988
16,500	10.0	0.066	108.9	7.0			530
70,000	10.0	0.066	462.0	7.0			2,249
100,000	10.0	0.033	330.0	7.0			1,607

A.7 Portfolio Optimization Strategy

We discuss a strategy of optimizing the megafund portfolio via the techniques of American options pricing and dynamic programming[1]. Consider a biomedical portfolio of n drug candidates and let $Y_1, ..., Y_n$ denote the binary drug approval outcomes. We assume that the binary outcomes have a uniform marginal PoS $\mathbb{P}(Y_i = 1) = \pi$ and uniform Pearson's correlation $\text{Corr}(Y_i, Y_j) = \rho$. Let FV denote the future value of the drug candidate upon FDA approval, C the cost of clinical trial, T the duration of clinical trial and r the annual discount rate.

The key insight of this optimization strategy is to view the portfolio not as one simultaneous investment but rather a series of correlated real options, where the drug developer can choose to develop a subset of the drug candidates and terminate future developments if the previous drug candidates have failed. This strategy generates financial value by utilizing the correlation to generate better forecasts of future drug development outcomes based on historical outcomes.

To set up the dynamic programming algorithm, we use $V[k, m]$ to denote the net present value (including the real options) of the $n-m$ remaining untested drugs given k approvals out of the first m clinical trials. At each stage $t \in [n]$, the drug developer decides whether to terminate future developments based on the outcomes of previous $t-1$ drugs. Therefore, the recursive Bellman equation reads

$$V[k, m] = \Big[\frac{\pi_{k,m}(FV + V[k+1, m+1]) + (1 - \pi_{k,m})V[k, m+1]}{(1+r)^T} - C\Big]^+ \tag{A.10}$$

where the function $[x]^+ = \max\{x, 0\}$ and $\pi_{k,m}$ denotes the conditional probability

$$\pi_{k,m} = \mathbb{P}\Big(Y_{m+1} = 1 | \sum_{i=1}^{m} Y_i = k\Big) \tag{A.11}$$

We assume that the correlation structure is known so that $\pi_{k,m}$ can be computed to good accuracy before plugging in the Bellman equation A.11. The optimal solution

[1]We thank Leonid Kogan for suggesting this approach.

to the Bellman equation yields an investment strategy. The drug developer should continue to invest after seeing k drug approvals out of the first m trials if $V[k, m] > 0$ and vice versa. In addition, the value $V[0, 0]$ of the optimal solution is the net present value of the optimized portfolio.

Preliminary analysis shows that under a given set of financial parameters (n, π, FV, C, T, r), the optimal portfolio value $V[0, 0]$ increases with the correlation ρ between clinical trial outcomes. Ongoing research work aims to apply this framework to the realistic megafund portfolios such as the GBM megafund.

Appendix B

Supplements to Chapter 4

Table B.1: Baseline and alternative parameter values used in the five-factor analysis.

Parameter	Description	Baseline Value	Alternative Values
a	Incubation rate (per week) [76]	1	0.5, 0.75, 1.25, 1.5
κ	Weekly subject enrollment in each arm of RCT (per week)	100	50, 75, 125, 150
p_0^{nv}	Prior probability of having an ineffective non-vaccine anti-infective therapy [2]	77%	54.90%, 65.45%, 88.55%, 96.25%
ρ	Signal-to-noise ratio of treatment effect [72]	0.25	0.125, 0.1875, 0.3125, 0.375
Δt	Time needed to assess the efficacy of the treatment (week)	1	0, 0.5, 1.5, 2

Table B.2: Optimal sample size and Type I error rate α for Bayesian non-adaptive RCT on anti-infective therapeutics with R_0 close to 1

Disease[1]	R_0	I_0(%)	Sample Size	α^* (%)	Power (%)
COVID-19	1.25	0.1	185	13.1	90
	1.5	0.1	239	7.3	90
	1.75	0.1	250	6.5	90
	2	0.1	242	7.1	90
	4	0.1	158	17.3	90
	1.25	0.01	233	7.8	90
	1.5	0.01	340	2.4	90
	1.75	0.01	395	1.3	90
	2	0.01	399	1.2	90
	4	0.01	274	5.0	90
SARS	1.25	0.1	69	42.6	90
	1.5	0.1	140	20.9	90
	1.75	0.1	162	16.6	90
	2	0.1	164	16.3	90
	4	0.1	112	27.8	90
MERS	1.25	0.1	6	80.2	90
	1.5	0.1	51	50.8	90
	1.75	0.1	80	38.2	90
	2	0.1	63	45.2	90
	4	0.1	60	46.5	90

[1] μ denotes the disease morality, and I_0 the proportion of initial infected subjects. Sample size denotes the number of subjects enrolled in each arm of the RCT.

Table B.3: Simulation results of a Bayesian adaptive RCT on non-vaccine anti-infective therapeutics.

Epidemic Parameters[2]				Nonadaptive		Adaptive sample size H_0		Adaptive sample size H_1			
R_0	μ	I_0 (%)	n^*	α^* (%)	Power (%)	Mean (SD)	Median (IQR)	Mean (SD)	Median (IQR)	α (%)	Power (%)
2	COVID-19	0.1	281	4.7	90	141 (113)	107 (63, 182)	175 (121)	142 (90,226)	3.7	91.3
4	COVID-19	0.1	176	14.4	90	119 (86)	96 (59,153)	108 (85)	83 (48,139)	11.9	92.3
$R_0(t)$	COVID-19	0.1	213	9.7	90	129 (95)	101 (61, 168)	131 (97)	103 (63,168)	8.3	91.7
2	COVID-19	0.01	433	0.8	90	150 (128)	109 (64, 191)	272 (166)	223 (156,348)	0.6	91.1
4	COVID-19	0.01	290	4.2	90	143 (113)	108 (63, 185)	177 (119)	145 (92,229)	3.3	91.1
$R_0(t)$	COVID-19	0.01	345	2.3	90	146 (118)	111 (65, 188)	217 (119)	181 (117,282)	1.9	91.0
2	SARS	0.1	262	5.7	90	138 (107)	107 (63, 179)	161 (115)	130 (81,207)	4.6	91.4
4	SARS	0.1	167	15.8	90	118 (86)	94 (57, 154)	102 (82)	77 (46,133)	13.6	92.3
$R_0(t)$	SARS	0.1	194	11.9	90	126 (93)	100 (60, 165)	118 (90)	92 (55,154)	9.7	92.5
2	MERS	0.1	227	8.4	90	130 (97)	102 (62, 167)	140 (101)	112 (68,181)	7.4	91.6
4	MERS	0.1	149	19.0	90	110 (78)	89 (55, 142)	93 (78)	69 (39,122)	16.1	92.7
$R_0(t)$	MERS	0.1	163	16.5	90	116 (84)	93 (56, 151)	101 (82)	78 (45,121)	13.7	92.2

[2]Results are obtained from 10,000 Monte Carlo runs and assuming $L_D = 10$. R_0 denotes the basic reproduction number, μ the disease morality, and I_0 the proportion of initial infected subjects. Sample size refers to the number of subjects enrolled in each arm of the RCT. SD denotes standard deviation, and IQR the interquartile range about the median. $R_0(t)$ denotes the dynamic transmission model with $t_2 = 3$ weeks, $\tau = 1$ week, and $R_0(t)$ decreasing from 3 to 1.5 as time t increases.

Table B.4: Simulation results of a Bayesian adaptive RCT on vaccines.

Epidemic Parameters[3]				Nonadaptive		Adaptive sample size H_0		Adaptive sample size H_1			
R_0	μ	I_0 (%)	n^*	α^* (%)	Power (%)	Mean (SD)	Median (IQR)	Mean (SD)	Median (IQR)	α (%)	Power (%)
2	COVID-19	0.1	221	8.9	90	130 (98)	103 (62, 167)	136 (100)	109 (66,176)	7.5	91.9
4	COVID-19	0.1	129	23.4	90	105 (78)	84 (51,135)	80 (70)	58 (33,105)	19.2	92.2
$R_0(t)$	COVID-19	0.1	151	18.7	90	114 (83)	92 (56, 149)	94 (78)	69 (40,123)	15.4	92.4
2	COVID-19	0.01	377	1.6	90	148 (126)	110 (64, 187)	236 (150)	200 (130,300)	1.2	90.8
4	COVID-19	0.01	247	6.7	90	135 (105)	103 (63, 175)	150 (106)	121 (75,196)	5.4	91.3
$R_0(t)$	COVID-19	0.01	281	4.6	90	139 (109)	107 (62, 181)	170 (116)	139 (88,220)	3.7	91.4
2	SARS	0.1	201	11.0	90	127 (94)	100 (61, 165)	122 (91)	96 (58,159)	9.9	91.9
4	SARS	0.1	120	25.6	90	101 (74)	81 (49, 131)	76 (68)	55 (30,100)	20.9	93.4
$R_0(t)$	SARS	0.1	134	22.2	90	106 (76)	86 (52, 139)	83 (72)	60 (33,108)	18.3	92.5
2	MERS	0.1	166	16.0	90	116 (84)	92 (56, 152)	102 (84)	77 (45,132)	13.5	92.2
4	MERS	0.1	103	30.4	90	93 (70)	75 (44, 122)	67 (63)	46 (24,88)	25.7	93.5
$R_0(t)$	MERS	0.1	105	29.8	90	96 (71)	77 (45, 125)	68 (62)	48 (25,91)	25.4	93.3

[3]Results are obtained from 10,000 Monte Carlo runs and assuming $L_D = 10$. R_0 denotes the basic reproduction number, μ the disease morality, and I_0 the proportion of initial infected subjects. Sample size refers to the number of subjects enrolled in each arm of the RCT. SD denotes standard deviation, and IQR the interquartile range about the median. $R_0(t)$ denotes the dynamic transmission model with $t_2 = 3$ weeks, $\tau = 1$ week, and $R_0(t)$ decreasing from 3 to 1.5 as time t increases.

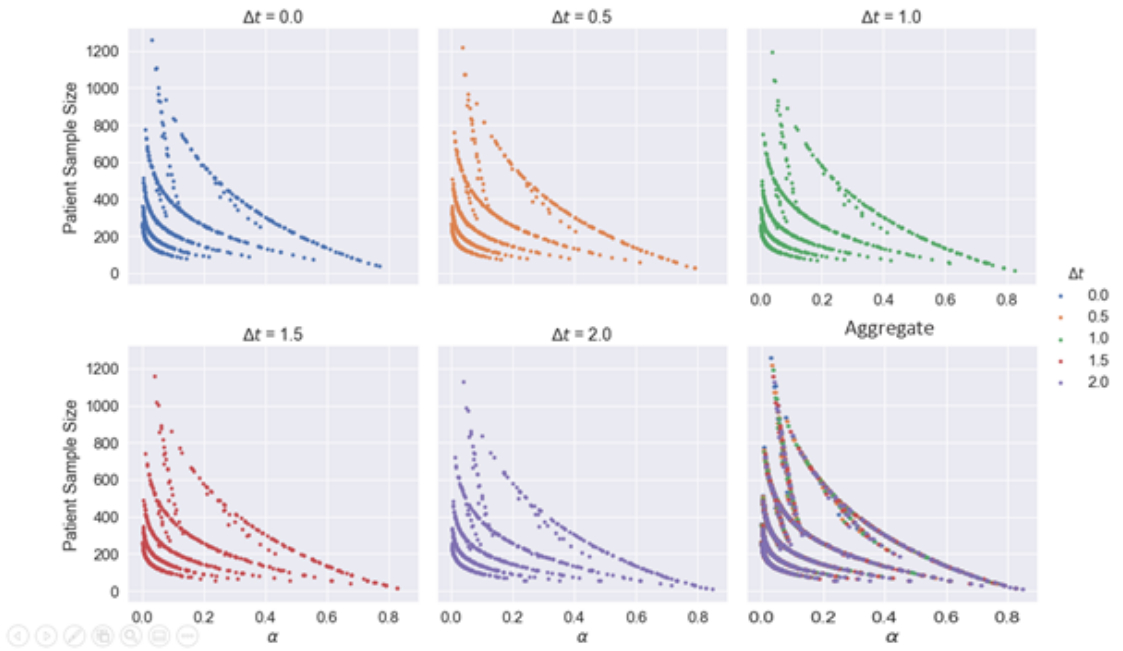


Figure B-1: Scatter plot of optimal Type I error rate α vs. sample size for different values of Δt, the time needed to assess the treatment efficacy (week). .

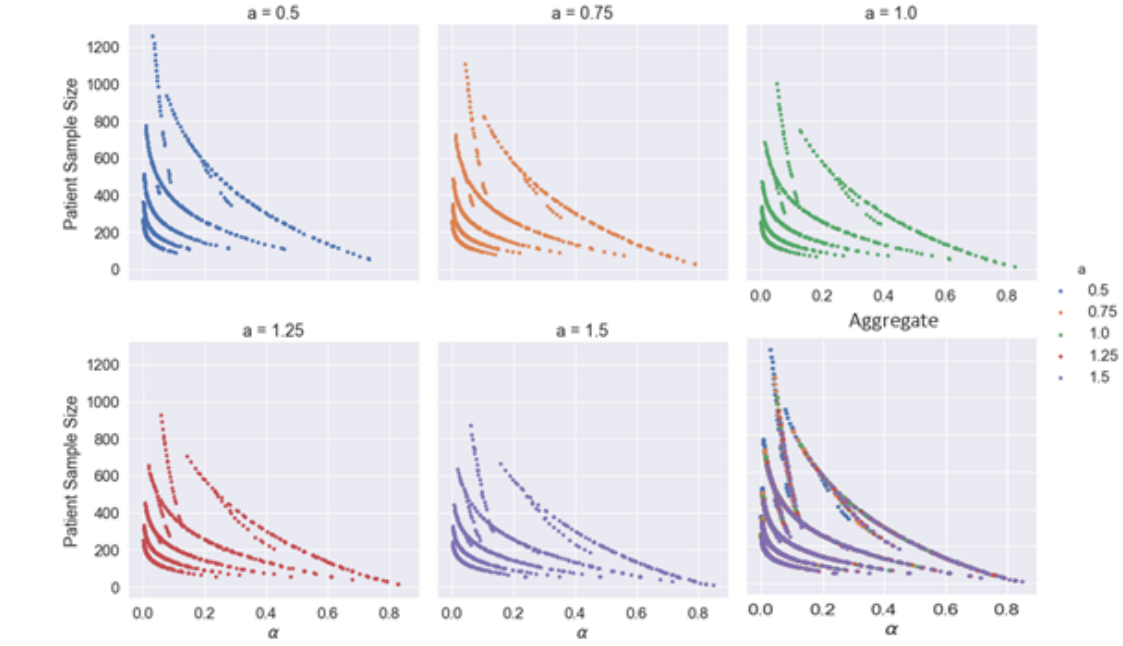


Figure B-2: Scatter plot of optimal Type I error rate α vs. sample size for different values of a, the incubation period (week) of the disease.

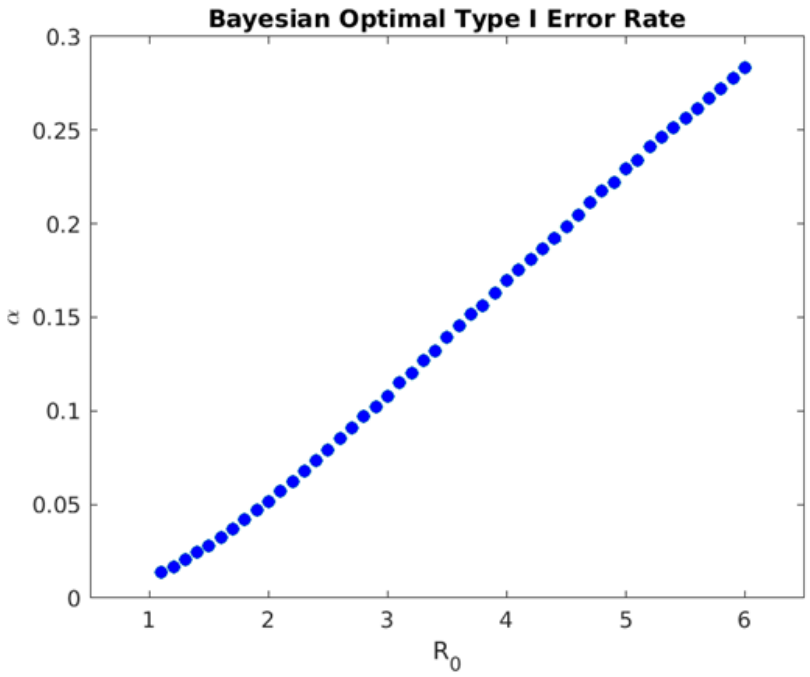


Figure B-3: Optimal Type I error rate α of non-adaptive Bayesian RCT monotonically increases with the basic reproduction number R_0 if we define the loss of making a Type I error as the absolute risk of being susceptible $S(t)NL_S$.

Assume $I_0 = 0.1\%, L_D = 100$ and disease mortality of COVID-19. This alternative definition is not very realistic. For an epidemic with $R_0 < 2$, the loss of Type I error converges to a large positive value $S(t)NL_S$ as time approaches the end of the epidemic outbreak. However, at the end of the outbreak, there are no more infected patients and thus no susceptible subjects. Therefore, the loss of Type I error should approach zero as $t \to T$. This is the case for the excess risk of susceptibility $(S(t) - S(T))NL_S$ but not for the absolute risk of susceptibility $S(t)NL_S$.

Appendix C

Supplements to Chapter 5

C.1 Computing Bayesian optimal sample size and Type I error

We review the procedure to compute the Bayesian optimal sample size and Type I error rate outlined in [14], where the authors assume a two-arm balanced randomized trial design. We extend this framework to a two-arm imbalanced trial design to model the phase 2 study of AMX0035. Let n_1 and n_2 denote the number of patients who are randomly assigned to the treatment and control arms, respectively. We assume that the primary treatment outcome (measured by the ALSFRS-R score) of each patient in the treatment arm $T_i \sim N(\mu_t, \sigma)$ follows a normal distribution with mean μ_t and standard deviation σ, and is independent and identically distributed (IID). Similarly, we assume that the outcome of patients in the control arm $C_i \sim N(\mu_c, \sigma)$ is also IID. We define the treatment effect δ as $\delta = \mu_t - \mu_c$.

We test the null hypothesis that there is no treatment effect, $H_0 : \delta = 0$ and the alternative hypothesis $H_0 : \delta = \delta_0 > 0$. To obtain a dimensionless measure of the treatment effect, we define the signal-to-noise ratio of treatment effect $\rho := \delta/\sigma$. Given the observed outcomes, we form the Wald statistic Z_n by

$$Z_n = \frac{\nu}{\sigma} \left(\overline{T} - \overline{C}\right) \tag{C.1}$$

where $\nu = \sqrt{n_1 n_2/(n_1+n_2)}$. The Wald statistic Z_n is a normal random variable with mean $\rho\nu$ and unit variance. The FDA performs a one-sided test and approves the drug if $Z_n > \lambda$. It is simple to calculate the Type I and Type II errors associated with a given threshold λ

$$\begin{aligned} \alpha(\lambda) &= P(Z_n > \lambda | H_0) = 1 - \Phi(\lambda) = \Phi(-\lambda) \\ \beta(\lambda) &= P(Z_n \leq \lambda | H_1) = \Phi(\lambda - \rho\nu) \end{aligned} \tag{C.2}$$

The total expected cost of the Bayesian decision analysis is

$$C(\lambda) = p_0\Big(\alpha(\lambda)C_{10} + (1-\alpha(\lambda))C_{00}\Big) + p_1\Big(\beta(\lambda)C_{01} + (1-\beta(\lambda))C_{11}\Big) \tag{C.3}$$

We use C_{ij} to denote the cost of choosing $\hat{H} = H_i$ given the reality $H = H_j$. Evaluating the first-order condition on $C(\lambda)$, we have that

$$C'(\lambda) = p_0\alpha'(\lambda)\big(C_{10} - C_{00}\big) + p_1\beta'(\lambda)\big(C_{01} - C_{11}\big) = 0 \quad \text{at} \quad \lambda = \lambda^* \tag{C.4}$$

Substituting Equations S2 and S3 for $\alpha(\lambda)$ and $\beta(\lambda)$, we get the optimal unconstrained decision threshold

$$\lambda^* = \frac{\rho\nu}{2} - \frac{\log\eta}{\rho\nu} \tag{C.5}$$

where $\eta = p_1(C_{01} - C_{11})/p_0(C_{10} - C_{00})$.

We also impose the constraint on the minimum Type II error $\beta(\lambda) = \Phi(\lambda - \rho\nu) \geq \beta_{min}$, which yields the optimal constrained decision threshold used in the simulations

$$\lambda^*(n_1, n_2) = \max\Big[\frac{\rho\nu}{2} - \frac{\log\eta}{\rho\nu}, \quad \rho\nu + \Phi^{-1}(\beta_{min})\Big] \tag{C.6}$$

To find the optimal sample size for the imbalanced trial with randomization ratio

2:1, we compute the optimal threshold $\lambda^*(n_1 = 2n, n_2 = n)$ for n from 1 to $n_{max} = 5000$, and report the optimal sample size n^* and the associated Type I and Type II error rates α^* and β^*.

Appendix D

Supplements to Chapter 6

D.1 Financial value calculation

We present the details of financial value calculation discussed in Section 6.4.4. The assumed values of the costs of capital are based on the work of [6], who uses the Capital Asset Pricing Model (CAPM) to estimate the cost of capital for pharmaceutical companies (8% to 10%) and biotechnology firms (10% to 15%) (see Table 4.5 in [6]). In the net present value (NPV) calculation, we assume that the phase 3 trial costs $100 million and takes 5 years to complete. If approved by the FDA, the drug will generate an annual profit of $2 billion over a 10-year period of market exclusivity. To discount the future cash flows, we define the initial time $t = 0$ as the beginning of phase 3 trial. If the drug is approved in year 5, we first discount the cash flows generated by its revenues to a future value $FV = 12.3$ billion (measured in year 5 currency) via

$$FV = \sum_{i=1}^{10} \frac{\$2 \text{ billion}}{(1 + 10\%)^i} = \$12.3 \text{ billion (in year 5 currency)} \tag{D.1}$$

We use the cost of capital 10% at the lower end of the estimate [6] in Equation D.1 to reflect the lower uncertainty in revenues of an approved drug. Next, we discount FV to year 0 and subtract the cost of phase 3 trial $C = \$100$ million to get the NPV in year 0:

$$NPV = \frac{FV}{(1+15\%)^5} - C = \$6 \text{ billion (in year 0 currency)} \tag{D.2}$$

Here we use the cost of capital 15% at the higher end of [6] in Equation D.2 to reflect the larger uncertainty in the drug approval outcome.

Table D.1: Drug and clinical trial features extracted from Informa database and used to predict the drug development outcome.

Feature	Examples	Type
Drug-Indication Pair		
Development outcome	Approved; Failed	Binary
Year of approval outcome	2004 to 2020	Numerical
Prior approval for another indication	True; False	Binary
Drug origin	Chemical, synthetic; Biological, protein, recombinant	Multi-label
Drug delivery medium	Capsule, hard; Capsule, soft; Powder; Solution; Suspension; Tablet	Multi-label
Drug delivery route	Oral; Inhaled; Topical; Injectable	Multi-label
Biological target	Cytokine/Growth factor; Ion channel; Enzyme; Receptor; Transporter	Multi-label
Pharmacology	Immunosuppressant; DNA inhibitor; Immunostimulant	Multi-label
Last year of phase 2 trials	1999 to 2019	Numerical
Clinical Trial		
Clinical trial status	Completed; Terminated	Binary
Trial attribute	Biomarker/Efficacy; Patient Preselection/Stratification	Multi-label
Target accrual	Positive integer	Numerical
Actual accrual	Positive integer	Numerical
Termination reason	Terminated, lack of efficacy; Competed, positive outcome/primary endpoint(s) met;	Multi-label
Therapeutic area	Oncology; Metabolic; Cardiovascular; Central Nervous System	Multi-label
Clinical trial duration (year)	Positive integer	Numerical
Design keyword	Randomized; Efficacy; Safety; Placebo control; Pharmacokinetics; Pharmacodynamics	Multi-label
Sponsor type	Industry, all other pharma; Government; Academic; Industry, top 20 pharma	Multi-label
Sponsor track record	Numbers of prior approved and failed drug-indication pairs	Numerical
Investigator track record	Same as in sponsor track record	Numerical
Number of clinical trial sites	Positive integer	Numerical
Clinical trial location	Canada; China; Japan; United States	Multi-label

Table D.2: Percentage of missing features in the P2APP dataset.

Feature	Total (2004 – 2020)	Training Set (2004 – 2018)	Testing Set (2019 – 2020)
Trial attribute	67.6	56.2	11.4
Termination reason	42.1	35.2	6.9
Drug delivery medium	32.2	28.4	3.8
Biological target	27.8	22.2	5.6
Target accrual	15.8	14.2	1.6
Clinical trial duration	10.3	9.3	1.0
Design keyword	7.9	6.8	1.1
Actual accrual	7.2	6.1	1.2
Pharmacology	3.4	2.8	0.6
Drug delivery route	0.7	0.6	0.1
Prior approval for another indication	0.4	0.4	0.0
Location	0.2	0.2	0.0

Table D.3: Model configuration and hyperparameter values of DB-VAE.

Hyperparameter	Value
Dimension of input features (m)	118
Dimension of latent space (d)	10
Number of neurons in each layer of DB-VAE encoder (ENC)	$[100, 50]$
Number of neurons in each layer of DB-VAE decoder (DEC)	$[50, 100]$
Number of training epochs	200
Learning rate of stochastic gradient descent	10^{-5}
Dropout rate	0.2
Training batch size	32
Weight of classification loss (λ_1)	10
Weight of categorical reconstruction loss (λ_2)	1
Weight of continuous reconstruction loss (λ_3)	1
Weight of KL divergence regularization (λ_4)	0.001
Imputing missing features	5-nearest-neighbor [120]
Minimum variance of each input feature (before imputation)	0.2

Table D.4: Sensitivity analysis of F_1 score against model hyperparameters.

F_1 score	λ_1	λ_2	λ_3	λ_4	Enc1	Enc2	d
0.48	10	1	1	0.001	100	50	10
0.49	21	17	18	0.055	92	48	10
0.45	29	27	12	0.079	88	58	12
0.47	3	2	1	0.083	103	59	16
0.45	14	24	23	0.012	67	58	14
0.46	12	16	8	0.077	88	40	11
0.48	18	30	22	0.023	60	40	14
0.48	18	17	11	0.003	92	53	15
0.50	21	17	15	0.063	92	49	8
0.46	19	16	21	0.014	90	47	9
0.47	23	18	17	0.045	92	51	8
0.47	21	18	16	0.044	94	51	13
0.46	19	16	17	0.086	92	50	11
0.48	25	14	18	0.083	95	47	9
0.44	22	20	15	0.035	90	50	11

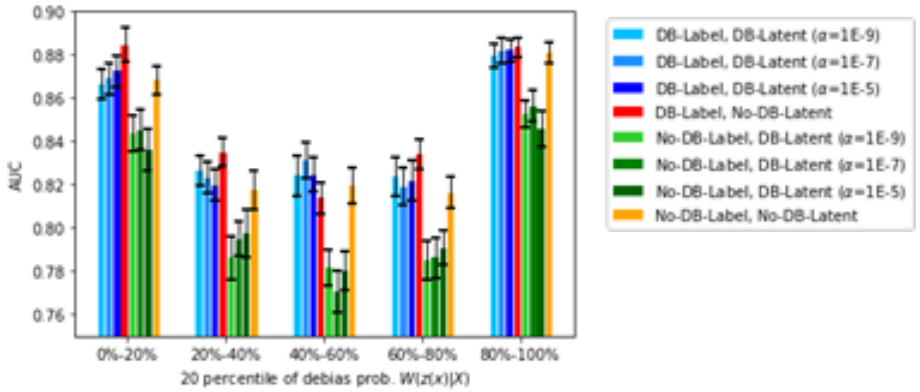


Figure D-1: AUC of DB-VAE models evaluated on the test dataset (2019-2020).

The un-debiased model (No-DB-Label, No-DB-Latent, plotted in orange) achieves high AUC in each quintile of $W(z(x)|X_{test})$, despite having the lowest true positive rate and F1 score (Table 6.3). High AUC in binary classification with imbalanced class labels can be misleading since a small number of correct predictions on the minority class (i.e., approved drugs) may have a disproportional impact on increasing the AUC [121]. Therefore, we do not use the AUC as a performance metric for the DB-VAE.

Appendix E

Supplements to Chapter 7

Table E.1: Clinical trial features extracted from the Informa database and used to predict the trial duration.

Feature	Examples	Type
Trial duration (year)	Positive real number	Numerical
Drug origin	Chemical, synthetic; Biological, protein, antibody	Multi-label
Drug delivery medium	Capsule, hard; Capsule, soft; Powder; Solution; Suspension; Tablet	Multi-label
Drug delivery route	Oral; Inhaled; Topical; Injectable	Multi-label
Clinical trial status	Completed; Terminated	Binary
Target accrual	Positive integer	Numerical
Percentage of target accrual	Positive real number	Numerical
Indication group	Oncology; Metabolic; Cardiovascular; Neurological; Infectious Disease	Multi-label
Sponsor type	Industry, top 20 pharma; Industry, all other pharma; Academic; Government; Coorperative group	Multi-label
Clinical trial location (aggregated in continents)	North America; Europe; Asia	Multi-label

Appendix F

Supplements to Chapter 8

F.1 Standard errors of ICC estimators

We provide the exact finite-sample standard errors of ICC estimators [45], [46].

Estimator 1: ANOVA estimator

$$\hat{\rho}_A = \frac{\mathrm{MS}_B - \mathrm{MS}_W}{\mathrm{MS}_B + (n_A - 1)\mathrm{MS}_W} \tag{F.1}$$

its finite-sample variance

$$\begin{aligned}
\mathrm{Var}(\hat{\rho}_A) = \frac{\big(n_A(K-1)(N-K)N\big)^2}{\lambda^4} \times \Big\{ & 2K + \Big(\frac{1}{\pi(1-\pi)} - 6\Big)\sum_{k=1}^{K}\frac{1}{N_k} \\
& + \rho\Big[\Big(\frac{1}{\pi(1-\pi)} - 6\Big)\sum_{k=1}^{K}\frac{1}{N_k} - 2N + 7K - \frac{8K^2}{N} - \frac{2K(1-K/N)}{\pi(1-\pi)} \\
& \quad + \Big(\frac{1}{\pi(1-\pi)} - 3\Big)\sum_{k=1}^{K} N_k^2\Big] \\
& + \rho^2\Big[\frac{N^2-K^2}{\pi(1-\pi)} - 2N - K + \frac{4K^2}{N} + \Big(7 - \frac{8K}{N} - \frac{2(1-K/N)}{\pi(1-\pi)}\Big)\sum_{k=1}^{K} N_k^2\Big] \\
& + \rho^3\Big[\Big(\frac{1}{\pi(1-\pi)} - 4\Big)\Big(\frac{N-K}{N}\Big)^2\Big(\sum_{k=1}^{K} N_k^2 - N\Big)\Big]\Big\}
\end{aligned} \tag{F.2}$$

and the auxiliary variables

$$
\begin{aligned}
n_A &= \frac{1}{K-1}\Big(N - \frac{\sum_{k=1}^{K} n_k^2}{N}\Big) \\
\mathrm{MS}_B &= \frac{1}{K-1}\Big(\sum_{k=1}^{K}\frac{S_k^2}{N_k} - \frac{(\sum_{k=1}^{K} S_k)^2}{N}\Big) \\
\mathrm{MS}_W &= \frac{1}{N-K}\Big(\sum_{k=1}^{K} S_k - \sum_{k=1}^{K}\frac{S_k^2}{N_k}\Big) \\
\lambda &= (N-K)\big(N-1-(K-1)n_A\big)\rho + N(K-1)(n_A-1)
\end{aligned}
\tag{F.3}
$$

Estimator 2: Fleiss-Cuzick (FC) estimator

$$
\hat{\rho}_{FC} = 1 - \frac{1}{(N-K)\hat{\pi}(1-\hat{\pi})}\sum_{k=1}^{K}\frac{S_k(n_k - S_k)}{n_k} \tag{F.4}
$$

its finite-sample variance

$$
\begin{aligned}
&\mathrm{Var}(\hat{\rho}_{FC}) = (1-\rho)\times \\
&\quad\Big\{\Big(\frac{1}{\pi(1-\pi)} - 6\Big)\frac{1}{(N-K)^2}\sum_{k=1}^{K}\frac{1}{N_k} + \Big(2N + 4K - \frac{K}{\pi(1-\pi)}\Big)\frac{K}{N(N-K)^2} \\
&\quad + \rho\Big[\Big(\frac{\sum_{k=1}^{K} N_k^2}{N^2\pi(1-\pi)} - \frac{(3N-2K)(N-2K)\sum_{k=1}^{K} N_k^2}{N^2(N-K)^2} - \frac{2N-K}{(N-K)^2}\Big] \\
&\quad + \rho^2\Big[\Big(4 - \frac{1}{\pi(1-\pi)}\Big)\frac{\sum_{k=1}^{K} N_k^2 - N}{N^2}\Big]\Big\}
\end{aligned}
\tag{F.5}
$$

and the auxiliary variable

$$
\hat{\pi} = \frac{\sum_{k=1}^{K} S_k}{N} \tag{F.6}
$$

Estimator 3: Pearson estimator

$$\hat{\rho}_P = \frac{1}{\hat{\mu}(1-\hat{\mu})}\Big(\frac{\sum_{k=1}^{K} S_k(S_k-1)}{\sum_{k=1}^{K} N_k(N_k-1)} - \hat{\mu}^2\Big) \tag{F.7}$$

its finite-sample variance

$$\begin{aligned}\mathrm{Var}(\hat{\rho}_P) = \frac{1-\rho}{[\sum_{k=1}^{K} N_k(N_k-1)]^2} \times \Big\{ & 2\sum_{k=1}^{K} N_k(N_k-1) \\ & + \rho\Big[\Big(\frac{1}{\pi(1-\pi)} - 3\Big)\sum_{k=1}^{K} N_k^2(N_k-1)^2\Big] \\ & + \rho^2\Big[\Big(4 - \frac{1}{\pi(1-\pi)}\Big)\sum_{k=1}^{K} N_k(N_k-1)^3\Big]\Big\}\end{aligned} \tag{F.8}$$

and the auxiliary variable

$$\hat{\mu} = \frac{\sum_{k=1}^{K} S_k(N_k-1)}{\sum_{k=1}^{K} N_k(N_k-1)} \tag{F.9}$$

Appendix G

Supplements to Chapter 10

G.1 Summary of ethical debate of COVID-19 HCTs

We summarize the six major aspects of the ethical debate surrounding COVID-19 HCTs in the bioethics literature.

Aspect 1: social value

- **Justification**. Accelerate vaccine development, thereby ending the pandemic sooner and avoiding large numbers of infections and deaths. Rapidly identify and control the harms of vaccine-enhanced disease by testing on a small group of participants. [34], [83], [175]–[177], [200], [205]

- **Critique**. HCT requires much more preparation (e.g., identifying and manufacturing low-risk virus strains, identifying low-risk populations, conducting dose-escalation studies, and establishing HCT protocol with input from regulators and local community of HCT site). Some experts estimate it would take 1 to 2 years to set up the HCT, which makes an HCT unlikely to accelerate vaccine development [197], [198], [200].

- **Critique**. Death or severe illness in HCT undermines stakeholder's confidence and public trust in vaccine development [83], [197], [201].

- **Critique**. Social value of HCT is overstated since equitable access to vaccines cannot be ensured, especially in low- and middle-income regions [199].

Aspect 2: scientific value

- **Justification**. Test multiple vaccine candidates in parallel and optimize vaccine dosage for safety and efficacy. Test on variants of COVID-19. Study correlates of protection, pathogenesis, disease progression, and transmission more precisely than standard vaccine trials. This helps inform public health policy (e.g., if booster shots are needed sometime after vaccination) [83], [175], [177], [198], [200], [205].

- **Critique**. There is an intrinsic tension between minimizing risks by selecting young, healthy participants and maximizing generalizability to high-risk groups who are older and with co-morbidities. Need to conduct follow-up field trials to test vaccine safety and efficacy on a much larger population to receive FDA approval. [175], [197], [198], [200], [201]

Aspect 3: feasibility and necessity

- **Justification**. Standard vaccine trials are infeasible during inter-pandemic periods or if public health policies suppress transmission so that few natural infections occur [200].

- **Justification**. The advent of effective vaccines makes testing new vaccines via standard trials difficult. However, trials for new vaccines should be prioritized to tackle new variants of concern and alleviate uneven access to vaccines. HCT is the suitable trial design [205].

- **Critique**. After January 2021, there are already several safe and highly effective vaccines for COVID-19 [202].

- **Critique**. There is no shortage of natural infection cases to test vaccine efficacy during the COVID-19 pandemic [197].

- **Critique**. If the number of COVID-19 cases can be kept low via nonpharmaceutical interventions, the urgency in vaccine development decreases [199].

- **Critique**. HCT diverts scarce medical resources in regions heavily afflicted by COVID-19, since participants receive additional medical care to minimize potential risks [197].

Aspect 4: risks and benefits to participants

- **Justification**. Potential direct benefits to participants (especially for those from communities with high infection rates and for essential workers such as healthcare providers) such as immunity by receiving effective vaccine and immunity by recovering from controlled infection [83], [177], [200].

- **Critique**. Highly uncertain and continuously evolving risks of COVID-19 to HCT participants and lack of rescue therapy to eliminate risks of severe infection and death. Unknown long-term risks of post-COVID conditions [83], [175]–[177], [200], [205].

- **Critique**. Informed consent does not provide sufficient protection to participants, who may be misled by "preventive misconception" – the false belief that participation in an HCT will provide some level of protection against COVID-19 [197], [201], [202].

Aspect 5: risks and benefits to third parties

- **Justification**. Potential indirect benefits to third parties at HCT sites since HCT participants are less likely to be infected and transmit the disease after recovery [83].

- **Critique**. Potential risk to third parties at HCT sites due to community spread of challenge virus and infection of HCT research staff [83], [198].

Aspect 6: fairness of site and participant selection

- **Justification**. HCT should only select participants who are at substantial risk of natural exposure to COVID-19 and face smaller marginal risk of participating in the HCT [177].

- **Justification**. Site selection of HCT is fairer than in a standard trial since the latter is conducted in populations with higher disease incidence and worse public health conditions [200].

- **Critique**. Socioeconomically disadvantaged populations are more heavily afflicted by COVID-19 and more likely to be attracted by monetary compensation in an HCT, raising concerns of potential exploitation and manipulation [199], [202].

Disease	Trial Identifier	Recruitment Status	Enrollment	Start-End	Sponsor
Cholera	NCT01895855	Completed	197	2013-2014	Emergent Biosolutions
	NCT03576183	Completed	34	2018-2018	Valneva Austria GmbH
Dengue	NCT00473135	Completed	60	2007-2010	NIAID
	NCT02021968	Completed	48	2013-2015	NIAID
	NCT02372175	Completed	27	2015-2019	USAMRDC
	NCT03416036	Completed	64	2017-2019	NIAID
	NCT03869060	Completed	9	2019-2019	SUNY - Upstate Medical University
Influenza	NCT02522754	Completed	174	2002-2004	Hvivo
	NCT01264601	Completed	31	2010-2012	University of Michigan
	NCT01971255	Completed	74	2013-2015	NIAID
	NCT02525055	Completed	46	2014-2014	Hvivo
	NCT04106817	Completed	35	2015-2015	WCCT Global
	NCT02594189	Completed	49	2015-2017	NIAID
	NCT02559505	Completed	134	2015-2020	University of Rochester
	NCT03180801	Completed	153	2016-2017	PepTCell Limited
	NCT02918006	Completed	179	2016-2018	Vaxart
	NCT04044352	Completed	76	2019-2020	NIAID
Malaria	NCT01556945	Completed	72	2001-2002	USAMRDC
	NCT00075049	Completed	104	2003-2006	USAMRDC
	NCT00385047	Completed	41	2006-2007	USAMRDC
	NCT00408369	Completed	19	2006-2007	Immtech Pharmaceuticals, Inc
	NCT00392015	Completed	59	2006-2020	USAMRDC
	NCT00761020	Completed	31	2008-2009	USAMRDC
	NCT00984256	Completed	35	2009-2010	USAMRDC
	NCT00935623	Completed	12	2009-2011	USAMRDC
	NCT00744133	Completed	37	2009-2011	NIAID
	NCT00870987	Completed	82	2009-2015	USAMRDC
	NCT01058226	Completed	6	2010-2010	Seattle Children's Research Institute
	NCT01157897	Completed	41	2010-2012	USAMRDC
	NCT01500980	Completed	36	2011-2012	Seattle Children's Research Institute
	NCT01397227	Completed	36	2011-2012	Crucell Holland BV
	NCT01441167	Completed	64	2011-2013	NIAID
	NCT01540474	Completed	36	2012-2012	USAMRDC
	NCT01546389	Completed	30	2012-2013	NIAID
	NCT02174978	Completed	39	2014-2015	USAMRDC
	NCT01994525	Completed	54	2014-2017	USAMRDC
	NCT04072302	Completed	86	2014-2017	Novartis Pharmaceuticals
	NCT02450578	Completed	22	2015-2016	Medicines for Malaria Venture
	NCT02780154	Completed	30	2016-2016	NIAID
	NCT02773979	Completed	28	2016-2018	NIAID
	NCT02661373	Completed	44	2016-2018	St. Jude Children's Research Hospital
	NCT03162614	Completed	154	2017-2018	GlaxoSmithKline
	NCT03168854	Completed	26	2017-2019	NIAID
	NCT03824236	Completed	61	2019-2019	GlaxoSmithKline
	NCT03882528	Completed	12	2019-2019	USAMRDC
	NCT03952650	Completed	393	2019-2021	NIAID
Norovirus	NCT02473224	Completed	44	2015-2018	NIAID
SARS-CoV-2	NCT04865237	Active, not recruiting	36	2021-2022	Imperial College London
	NCT04864548	Recruiting	64	2021-2022	University of Oxford
Tuberculosis	NCT02088892	Completed	52	2014-2015	University of Oxford
Typhoid	NCT01405521	Active, not recruiting	99	2011-2021	University of Oxford
	NCT02192008	Active, not recruiting	125	2014-2021	University of Oxford
	NCT03067961	Active, not recruiting	40	2017-2021	University of Oxford

Figure G-1: Select human challenge trials since 2000.

Bibliography

[1] J. W. Scannell, A. Blanckley, H. Bolden, and B. Warrington, "Diagnosing the decline in pharmaceutical R&D efficiency," *Nature Reviews Drug Discovery*, vol. 11, pp. 191–200, 2012. DOI: 10.1038/nrd3681.

[2] C. H. Wong, K. W. Siah, and A. W. Lo, "Estimation of clinical trial success rates and related parameters," *Biostatistics*, vol. 20, no. 2, pp. 273–286, 2019. DOI: 10.1093/biostatistics/kxx069.

[3] MIT Laboratory for Financial Engineering. "Estimates of Clinical Trial Probabilities of Success (PoS)." (2021), [Online]. Available: https://projectalpha.mit.edu/pos/ (visited on 11/08/2021).

[4] O. J. Wouters, M. McKee, and J. Luyten, "Estimated research and development investment needed to bring a new medicine to market, 2009-2018," *JAMA*, vol. 323, no. 8, pp. 844–853, 2020. DOI: 10.1001/jama.2020.1166.

[5] L. Martin, M. Hutchens, and C. Hawkins, "Clinical trial cycle times continue to increase despite industry efforts," *Nature Reviews Drug Discovery*, vol. 16, no. 157, 2017. DOI: 10.1038/nrd.2017.21.

[6] S. E. Harrington, "Cost of capital for pharmaceutical, biotechnology, and medical device firms," in *The Oxford Handbook of the Economics of the Biopharmaceutical Industry*, ser. The Oxford Handbook of the Economics of the Biopharmaceutical Industry, P. M. Danzon and S. Nicholson, Eds., Oxford University Press, 2012. DOI: 10.1093/oxfordhb/9780199742998.013.0004.

[7] D. Butler, "Translational research: Crossing the valley of death," *Nature*, vol. 453, pp. 840–842, 2008. DOI: 10.1038/453840a.

[8] A. W. Lo and R. T. Thakor, "Financing medical innovation," *Working paper*, 2021.

[9] J.-M. Fernandez, R. M. Stein, and A. W. Lo, "Commercializing biomedical research through securitization techniques," *Nature Biotechnology*, vol. 30, pp. 964–975, 2012. DOI: 10.1038/nbt.2374.

[10] D. E. Fagnan, A. A. Gromatzky, R. M. Stein, J.-M. Fernandez, and A. W. Lo, "Financing drug discovery for orphan diseases," *Drug Discovery Today*, vol. 19, no. 5, pp. 533–538, 2014. DOI: 10.1016/j.drudis.2013.11.009.

[11] A. W. Lo, C. Ho, J. Cummings, and K. S. Kosik, "Parallel Discovery of Alzheimer's Therapeutics," *Science Translational Medicine*, vol. 6, p. 241cm5, 241 2014. DOI: 10.1126/scitranslmed.3008228.

[12] N. B. T. Society. "Brain tumor investment^{TM}*fund*." (2021), [Online]. Available: https://braintumor.org/our-research/investment-fund/ (visited on 03/28/2022).

[13] B. Hauber, B. Mange, M. Zhou, *et al.*, "Parkinson's Patients' Tolerance for Risk and Willingness to Wait for Potential Benefits of Novel Neurostimulation Devices: A Patient-Centered Threshold Technique Study," *MDM policy & practice*, vol. 6, no. 1, p. 2 381 468 320 978 407, 2021. DOI: 10.1177/2381468320978407.

[14] L. Isakov, A. W. Lo, and V. Montazerhodjat, "Is the FDA too conservative or too aggressive?: A Bayesian decision analysis of clinical trial design," *Journal of Econometrics*, vol. 211, no. 1, pp. 117–136, 2019. DOI: 10.1016/j.jeconom.2018.12.009.

[15] V. Montazerhodjat, S. E. Chaudhuri, D. J. Sargent, and A. W. Lo, "Use of Bayesian Decision Analysis to Minimize Harm in Patient-Centered Randomized Clinical Trials in Oncology," *JAMA Oncology*, vol. 3, no. 9, e170123–e170123, 2017. DOI: 10.1001/jamaoncol.2017.0123.

[16] S. E. Chaudhuri, M. P. Ho, T. Irony, M. Sheldon, and A. W. Lo, "Patient-centered clinical trials," *Drug Discovery Today*, vol. 23, no. 2, pp. 395–401, 2018. DOI: https://doi.org/10.1016/j.drudis.2017.09.016.

[17] S. E. Chaudhuri and A. W. Lo, "Incorporating Patient Preferences via Bayesian Decision Analysis," *Clinical Journal of the American Society of Nephrology*, vol. 16, pp. 639–641, 2021. DOI: 10.2215/CJN.12110720.

[18] R. H. El-Maraghi and E. 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*, vol. 26, no. 8, pp. 1346–1354, 2008. DOI: 10.1200/JCO.2007.13.5913.

[19] L. Malik, A. Mejia, H. Parsons, *et al.*, "Predicting success in regulatory approval from Phase I results," *Cancer Chemotherapy and Pharmacology*, vol. 74, pp. 1099–1103, 2014. DOI: 10.1007/s00280-014-2596-4.

[20] J. DiMasi, J. Hermann, K. Twyman, *et al.*, "A Tool for Predicting Regulatory Approval After Phase II Testing of New Oncology Compounds," *Clinical Pharmacology & Therapeutics*, vol. 98, no. 5, pp. 506–513, 2015. DOI: 10.1002/cpt.194.

[21] A. W. Lo, K. W. Siah, and C. H. Wong, "Machine learning with statistical imputation for predicting drug approvals," *Harvard Data Science Review*, vol. 1, no. 1, 2019. DOI: 10.1162/99608f92.5c5f0525.

[22] Citeline. (2022), [Online]. Available: https://pharmaintelligence.informa.com/products-and-services/data-and-analysis/citeline (visited on 01/23/2022).

[23] A. W. Lo, K. W. Siah, and C. H. Wong, "Estimating probabilities of success of vaccine and other anti-infective therapeutic development programs," *Harvard Data Science Review*, May 14, 2020. DOI: 10.1162/99608f92.e0c150e8.

[24] K. W. Siah, N. W. Kelley, S. Ballerstedt, *et al.*, "Predicting drug approvals: The Novartis data science and artificial intelligence challenge," *Patterns*, vol. 2, no. 8, p. 100 312, 2021. DOI: 10.1016/j.patter.2021.100312.

[25] V. Montazerhodjat, D. M. Weinstock, and A. W. Lo, "Buying cures versus renting health: Financing health care with consumer loans," *Science Translational Medicine*, vol. 8, 327ps6, 327 2016. DOI: 10.1126/scitranslmed.aad6913.

[26] K. W. Siah, Q. Xu, K. Tanner, O. Futer, J. J. Frishkopf, and A. W. Lo, "Accelerating glioblastoma therapeutics via venture philanthropy," *Drug Discovery Today*, vol. 26, no. 7, pp. 1744–1749, 2021. DOI: 10.1016/j.drudis.2021.03.020.

[27] S. E. Chaudhuri, A. W. Lo, D. Xiao, and Q. Xu, "Bayesian Adaptive Clinical Trials for Anti-Infective Therapeutics During Epidemic Outbreaks," *Harvard Data Science Review*, 2020. DOI: 10.1162/99608f92.7656c213.

[28] Q. Xu, E. Ahmadi, A. Amini, D. Rus, and A. W. Lo, "Identifying and mitigating potential biases in predicting drug approvals," *Drug Safety*, 2022.

[29] Q. Xu and A. W. Lo, "Fair and responsible drug pricing: A cast study of Radius Health and *abaloparatide*," *Journal of Investment Management*, vol. 18, no. 1, pp. 90–98, 2020.

[30] S. E. Chaudhuri, K. Cheng, A. W. Lo, *et al.*, "A portfolio approach to accelerate therapeutic innovation in ovarian cancer," *Journal of Investment Management*, vol. 17, no. 2, pp. 5–16, 2019.

[31] S. Das, R. Rousseau, P. C. Adamson, and A. W. Lo, "New Business Models to Accelerate Innovation in Pediatric Oncology Therapeutics: A Review," *JAMA Oncology*, vol. 4, no. 9, pp. 1274–1280, 2018. DOI: 10.1001/jamaoncol.2018.1739.

[32] B. M. Alexander, S. Ba, and M. S. Berger, "Adaptive Global Innovative Learning Environment for Glioblastoma: GBM AGILE," *Clinical Cancer Research*, vol. 24, pp. 737–743, 2018. DOI: 10.1158/1078-0432.CCR-17-0764.

[33] J. T. Vu, B. K. Kaplan, S. Chaudhuri, M. K. Mansoura, and A. W. Lo, "Financing vaccines for global health security," *Journal of Investment Management*, vol. 20, no. 2, pp. 1–17, 2022.

[34] D. A. Berry, S. Berry, P. Hale, *et al.*, "A cost/benefit analysis of clinical trial designs for COVID-19 vaccine candidates," *PLOS ONE*, vol. 15, no. 12, pp. 1–17, 2020. DOI: 10.1371/journal.pone.0244418.

[35] L. R. Baden, C. G. Solomon, M. F. Greene, R. B. D'Agostino, and D. Harrington, "The fda and the importance of trust," *New England Journal of Medicine*, vol. 383, e148, 2020. DOI: 10.1056/NEJMe2030687.

[36] E. S. Pronker, T. C. Pronker, H. Commandeur, E. H. J. H. M. Claassen, and A. D. M. E. Osterhaus, "Risk in Vaccine Research and Development Quantified," *PLoS ONE*, vol. 8, no. 3, e57755, 2013. DOI: 10.1371/journal.pone.0057755.

[37] A. Pharmaceuticals. "Amylyx pharmaceuticals submits new drug application (nda) for amx0035 for the treatment of als." (2021), [Online]. Available: https://www.amylyx.com/media/amylyx-pharmaceuticals-submits-new-drug-application-nda-for-amx0035-for-the-treatment-of-als (visited on 11/02/2021).

[38] S. Paganoni and et al., "Trial of sodium phenylbutyrate–taurursodiol for amyotrophic lateral sclerosis," *New England Journal of Medicine*, vol. 373, pp. 919–930, 2020. DOI: 10.1056/NEJMoa1916945.

[39] ——, "Long-term survival of participants in the centaur trial of sodium phenylbutyrate-taurursodiol in amyotrophic lateral sclerosis," *Muscle & Nerve*, vol. 63, pp. 31–39, 2021. DOI: 10.1002/mus.27091.

[40] G. Beinse, V. Tellier, V. Charvet, *et al.*, "Prediction of drug approval after phase i clinical trials in oncology: Resolved2," *JCO Clinical Cancer Informatics*, vol. 3, pp. 1–10, 2019. DOI: 10.1200/CCI.19.00023.

[41] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, "A survey on bias and fairness in machine learning," *ACM Computing Surveys*, vol. 54, no. 6, 2021. DOI: 10.1145/3457607.

[42] A. Amini, A. P. Soleimany, W. Schwarting, S. N. Bhatia, and D. Rus, "Uncovering and mitigating algorithmic bias through learned latent structure," in *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, ser. AIES '19, Honolulu, HI, USA, 2019, pp. 289–295. DOI: 10.1145/3306618.3314243.

[43] P. Wang, Y. Li, and C. K. Reddy, "Machine Learning for Survival Analysis: A Survey," *ACM Computing Surveys*, vol. 51, no. 6, pp. 1–36, 2019. DOI: 10.1145/3214306.

[44] A. W. Lo and K. W. Siah, "Financing Correlated Drug Development Projects," *The Journal of Structured Finance*, 2020. DOI: 10.3905/jsf.2020.1.114.

[45] M. S. Ridout, C. G. B. Demktrio, and D. Firth, "Estimating intraclass correlation for binary data," *Biometrics*, vol. 55, pp. 137–148, 1999. DOI: 10.1111/j.0006-341x.1999.00137.x.

[46] S. Wu, C. M. Crespi, and W. K. Wong, "Comparison of methods for estimating the intraclass correlation coefficient for binary responses in cancer prevention cluster randomized trials," *Contemporary Clinical Trials*, vol. 33, pp. 869–880, 2012. DOI: 10.1016/j.cct.2012.05.004.

[47] K.-Y. Liang and S. L. Zeger, "Longitudinal data analysis using generalized linear models," *Biometrika*, vol. 73, no. 1, pp. 13–22, 1986. DOI: 10.2307/2336267.

[48] R. L. Prentice, "Correlated binary regression with covariates specific to each binary observation," *Biometrics*, vol. 44, no. 4, pp. 1033–1048, 1988. DOI: 10.2307/2531733.

[49] J. Yan and J. Fine, "Estimating equations for association structures," *Statistics in Medicine*, vol. 44, no. 4, pp. 1033–1048, 2004.

[50] Q. T. Ostrom, N. Patil, G. Cioffi, K. Waite, C. Kruchko, and J. S. Barnholtz-Sloan, "CBTRUS Statistical Report: Primary Brain and Other Central Nervous System Tumors Diagnosed in the United States in 2013–2017," *Neuro-Oncology*, vol. 22, no. Supplement 1, pp. iv1–iv96, Oct. 2020. DOI: 10.1093/neuonc/noaa200.

[51] M. E. Davis, "Glioblastoma: Overview of disease and treatment," *Clinical Cancer Research*, vol. 20, no. 5, S2–S8, 2016. DOI: 10.1188/16.CJON.S1.2-8.

[52] J. P. Thakkar, T. A. Dolecek, C. Horbinski, *et al.*, "Epidemiologic and molecular prognostic review of glioblastoma," *Cancer Epidemiology, Biomarkers Prevention*, vol. 23, no. 10, pp. 1985–1996, 2014. DOI: 10.1158/1055-9965.EPI-14-0275.

[53] G. C. for Adaptive Research. "A trial to evaluate multiple regimens in newly diagnosed and recurrent glioblastoma (gbm agile)." (2019), [Online]. Available: https://clinicaltrials.gov/ct2/show/NCT03970447 (visited on 05/31/2019).

[54] A. G. I. L. E. for Glioblastoma: GBM AGILE, *Clinical Cancer Research*, vol. 24, no. 4, pp. 737–743, 2018. DOI: 10.1158/1078-0432.CCR-17-0764.

[55] C. for Epidemic Preparedness Innovations. "Our portfolio." (2022), [Online]. Available: https://cepi.net/research_dev/our-portfolio/ (visited on 02/19/2022).

[56] N. Chaudhary, D. Weissman, and K. A. Whitehead, "Mrna vaccines for infectious diseases: Principles, delivery and clinical translation," *Nature Reviews Drug Discovery*, vol. 20, pp. 817–838, 2021. DOI: 10.1038/s41573-021-00283-5.

[57] D. Gouglas and K. Marsh, "Prioritizing investments in new vaccines against epidemic infectious diseases: A multi-criteria decision analysis," *Journal of Multi-Criteria Decision Analysis*, vol. 26, no. 3, pp. 153–163, 2019. DOI: 10.1002/mcda.1683.

[58] ——, "Prioritizing investments in rapid response vaccine technologies for emerging infections: A portfolio decision analysis," *PLOS ONE*, vol. 16, no. 2, e0246235, 2021. DOI: 10.1371/journal.pone.0246235.

[59] A. Ahuja, S. Athey, A. Baker, *et al.*, "Preparing for a pandemic: Accelerating vaccine availability," *AEA Papers and Proceedings*, vol. 111, pp. 331–335, 2021. DOI: 10.1257/pandp.20211103.

[60] S. Jain, A. Venkataraman, M. E. Wechsler, and N. A. Peppas, "Messenger rna-based vaccines: Past, present, and future directions in the context of the covid-19 pandemic," *Advanced Drug Delivery Reviews*, vol. 179, p. 114 000, 2021. DOI: 10.1016/j.addr.2021.114000.

[61] D. Gouglas, T. T. Le, K. Henderson, *et al.*, "Estimating the cost of vaccine development against epidemic infectious diseases: A cost minimisation study," *Lancet Global Health*, vol. 6, no. 12, pp. 1386–1396, 2018. DOI: 10.1016/S2214-109X(18)30346-2.

[62] H. Qi and D. Sun, "A quadratically convergent newton method for computing the nearest correlation matrix," *SIAM Journal on Matrix Analysis and Applications*, vol. 28, no. 2, pp. 360–385, 2006. DOI: 10.1137/050624509.

[63] Z. Kis, C. Kontoravdi, R. Shattock, and N. Shah, "Resources, production scales and time required for producing rna vaccines for the global pandemic demand," *Vaccines*, vol. 9, no. 1, 2021. DOI: 10.3390/vaccines9010003.

[64] Z. Kis and Z. Rizvi, "How to make enough vaccine for the world in one year?" Public Citizen, Technical Report, 2021.

[65] D. Jimenez. "Covid-19: Vaccine pricing varies wildly by country and company." (2021), [Online]. Available: https://www.pharmaceutical-technology.com/features/covid-19-vaccine-pricing-varies-country-company/ (visited on 02/20/2022).

[66] U. C. for Disease Control and Prevention. "Cdc vaccine price list." (2022), [Online]. Available: https://www.cdc.gov/vaccines/programs/vfc/awardees/vaccine-management/price-list/index.html (visited on 01/09/2022).

[67] D. A. Berry, "Adaptive clinical trials in oncology," *Nature Reviews Clinical Oncology*, vol. 9, no. 4, pp. 199–207, 2011. DOI: 10.1038/nrclinonc.2011.165.

[68] J. Woodcock and L. M. LaVange, "Master protocols to study multiple therapies, multiple diseases, or both," *New England Journal of Medicine*, vol. 377, pp. 62–70, 2017. DOI: 10.1056/NEJMra1510062.

[69] V. Montazerhodjat, J. J. Frishkopf, and A. W. Lo, "Financing drug discovery via dynamic leverage," *Drug Discovery Today*, vol. 21, no. 3, pp. 410–414, 2016. DOI: 10.1016/j.drudis.2015.12.004.

[70] U. F. D. Administration. "Vaccine product approval process." (2018), [Online]. Available: https://www.fda.gov/vaccines-blood-biologics/development-approval-process-cber/vaccine-product-approval-process (visited on 01/30/2018).

[71] E. S. Pronker, T. C. Weenen, H. Commandeur, E. H. J. H. M. Claassen, and A. D. M. E. Osterhaus, "Risk in vaccine research and development quantified," *PLoS One*, vol. 8, no. 3, e57755, 2013. DOI: 10.1371/journal.pone.0057755.

[72] S. E. Chaudhuri, M. P. Ho, T. Irony, M. Sheldon, and A. W. Lo, "Patient-centered clinical trials," *Drug Discovery Today*, vol. 23, no. 2, pp. 395–401, 2018. DOI: 10.1016/j.drudis.2017.09.016.

[73] S. E. Chaudhuri and A. W. Lo, "Bayesian adaptive patient-centered clinical trials," 2018.

[74] D. A. Berry, "The brave new world of clinical cancer research: Adaptive biomarker-driven trials integrating clinical practice with clinical research," *Molecular Oncology*, vol. 9, no. 5, pp. 951–959, 2015. DOI: 10.1016/j.molonc.2015.02.011.

[75] A. D. Barker, C. C. Sigman, G. J. Kelloff, N. M. Hylton, D. A. Berry, and E. L. J, "I-spy 2: An adaptive breast cancer trial design in the setting of neoadjuvant chemotherapy," *Clinical Pharmacology Therapeutics*, vol. 86, no. 1, pp. 97–100, 2009. DOI: 10.1038/clpt.2009.68.

[76] Z. Yang, Z. Zeng, K. Wang, S.-S. Wong, W. Liang, and M. Zanin, "Modified seir and ai prediction of the epidemics trend of covid-19 in china under public health interventions," *Journal of Thoracic Disease*, vol. 12, no. 3, pp. 165–174, 2020. DOI: 10.21037/jtd.2020.02.64.

[77] J. T. Wu, K. Leung, and G. M. Leung, "Nowcasting and forecasting the potential domestic and international spread of the 2019-ncov outbreak originating in wuhan, china: A modelling study," *Lancet*, vol. 395, no. 10225, pp. 689–697, 2020. DOI: 10.1016/S0140-6736(20)30260-9.

[78] Q. Li, X. Guan, P. Wu, *et al.*, "Early transmission dynamics in wuhan, china, of novel coronavirus–infected pneumonia," *New England Journal of Medicine*, vol. 382, no. 13, pp. 1199–1207, 2020. DOI: 10.1056/NEJMoa2001316.

[79] S. Zhao, Q. Lin, J. Ran, *et al.*, "Preliminary estimation of the basic reproduction number of novel coronavirus (2019-ncov) in china, from 2019 to 2020: A data-driven analysis in the early phase of the outbreak," *International Journal of Infectious Diseases*, vol. 92, pp. 214–217, 2020. DOI: 10.1016/j.ijid.2020.01.050.

[80] A. J. Kucharski, T. W. Russell, C. Diamond, *et al.*, "Early dynamics of transmission and control of covid-19: A mathematical modelling study," *The Lancet Infectious Diseases*, vol. 20, no. 5, pp. 553–558, 2020. DOI: 10.1016/S1473-3099(20)30144-4.

[81] G. Onder, G. Rezza, and S. Brusaferro, "Case-fatality rate and characteristics of patients dying in relation to covid-19 in italy," *JAMA*, vol. 323, no. 18, pp. 1775–1776, 2020. DOI: 10.1001/jama.2020.4683.

[82] P. G. T. Walker, C. Whittaker, O. J. Watson, *et al.*, "The impact of covid-19 and strategies for mitigation and suppression in low- and middle-income countries," *Science*, vol. 369, no. 6502, pp. 413–422, 2020. DOI: 10.1126/science.abc0035.

[83] W. H. Organization. "Coronavirus disease 2019 (covid-19) situation report–59." (2020), [Online]. Available: https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200319-sitrep-59-covid-19.pdf?sfvrsn=c3dcdef9_2 (visited on 03/19/2020).

[84] ——, "Consensus document on the epidemiology of severe acute respiratory syndrome (sars)." (2003), [Online]. Available: https://www.who.int/csr/sars/en/WHOconsensus.pdf (visited on 11/01/2003).

[85] ——, "Mers situation update, december 2019." (2019), [Online]. Available: http://www.emro.who.int/pandemic-epidemic-diseases/mers-cov/mers-situation-update-december-2019.html (visited on 12/01/2019).

[86] I. J. Onakpoya, C. J. Heneghan, and J. K. Aronson, "Post-marketing withdrawal of 462 medicinal products because of adverse drug reactions: A systematic review of the world literature," *BMC Medicine*, vol. 14, p. 10, 2016. DOI: 10.1186/s12916-016-0553-2.

[87] A. W. Lo, "Discussion: New directions for the FDA in the 21st century," *Biostatistics*, vol. 18, no. 3, pp. 404–407, 2017. DOI: 10.1093/biostatistics/kxx019.

[88] U. F. D. Administration. "Structured approach to benefit-risk assessment in drug regulatory decision-making: Draft pdufa v implementation plan—february 2013, fiscal years 2013–2017." (2013), [Online]. Available: https://www.fda.gov/media/84831/download (visited on 02/01/2013).

[89] R. H. Brown and A. Al-Chalabi, "Amyotrophic lateral sclerosis," *New England Journal of Medicine*, vol. 377, pp. 162–172, 2017. DOI: 10.1056/NEJMra1603471.

[90] P. Mehta, W. Kaye, J. Raymond, *et al.*, "Mmwr prevalence of amyotrophic lateral sclerosis – united states, 2015," *MMWR Morbidity and Mortality Weekly Report*, vol. 67, pp. 1285–1289, 2018. DOI: 10.15585/mmwr.mm6746a1.

[91] J. Larkindale, W. Yang, P. F. Hogan, *et al.*, "Cost of illness for neuromuscular diseases in the united states," *Muscle & Nerve*, vol. 49, pp. 431–438, 2014. DOI: 10.1002/mus.23942.

[92] U. F. D. Administration. "Update on amyotrophic lateral sclerosis (als) product development." (2021), [Online]. Available: https://www.fda.gov/vaccines-blood-biologics/cellular-gene-therapy-products/update-amyotrophic-lateral-sclerosis-als-product-development (visited on 03/02/2021).

[93] ClinicalTrials.gov. "Phase iii trial of amx0035 for amyotrophic lateral sclerosis treatment (phoenix)." (2021), [Online]. Available: https://clinicaltrials.gov/ct2/show/%20NCT05021536 (visited on 12/01/2021).

[94] A. Association. "Als association applauds amylyx's amx0035 announcement, urges swift fda approval." (2021), [Online]. Available: https://www.als.org/stories-news/als-association-applauds-amylyxs-amx0035-announcement-urges-swift-fda-approval (visited on 09/15/2021).

[95] U. F. D. Administration. "Guidance for industry: Fast track drug development programs – designation, development, and application review." (2006), [Online]. Available: https://permanent.fdlp.gov/LPS113471/UCM079736.pdf (visited on 01/01/2006).

[96] ——, "Guidance for industry: Expedited programs for serious conditions – drugs and biologics." (2014), [Online]. Available: https://www.fda.gov/media/86377/download (visited on 05/01/2014).

[97] M. Greener, "Drug safety on trial. last year's withdrawal of the anti-arthritis drug vioxx triggered a debate about how to better monitor drug safety even after approval," *EMBO Reports*, vol. 6, no. 3, pp. 202–204, 2005. DOI: 10.1038/sj.embor.7400353.

[98] J. K. Aronson, "Drug withdrawals because of adverse effects," in *A worldwide yearly survey of new data and trends in adverse drug reactions and interactions*, ser. Side Effects of Drugs Annual, J. Aronson, Ed., vol. 30, Elsevier, 2008, pp. xxxi–xxxv. DOI: 10.1016/S0378-6080(08)00064-0.

[99] R. McNaughton, G. Huet, and S. Shakir, "An investigation into drug products withdrawn from the eu market between 2002 and 2011 for safety reasons and the evidence used to support the decision-making," *BMJ Open*, vol. 4, e004221, 2014. DOI: 10.1136/bmjopen-2013-004221.

[100] ProCon.org. "35 fda-approved prescription drugs later pulled from the market." (2014), [Online]. Available: https://www.procon.org/35-prescription-drugs-pulled-from-the-market (visited on 02/07/2014).

[101] A. E. Elia, S. Lalli, M. R. Monsurrò, *et al.*, "Tauroursodeoxycholic acid in the treatment of patients with amyotrophic lateral sclerosis," *European Journal of Neurology*, vol. 23, no. 1, pp. 45–52, 2016. DOI: 10.1111/ene.12664.

[102] G. 2. M. N. D. Collaborators, "Global, regional, and national burden of motor neuron diseases 1990–2016: A systematic analysis for the global burden of disease study 2016," *Lancet Neurology*, vol. 17, no. 12, pp. 1083–1097, 2018. DOI: 10.1016/S1474-4422(18)30404-6.

[103] B. Freedman, "Equipoise and the ethics of clinical research," *New England Journal of Medicine*, vol. 317, pp. 141–145, 1987. DOI: 10.1056/NEJM198707163170304.

[104] J. Goffin, S. Baral, D. Tu, D. Nomikos, and L. Seymour, "Objective responses in patients with malignant melanoma or renal cell cancer in early clinical studies do not predict regulatory approval," *Clinical Cancer Research*, vol. 11, no. 16, pp. 5928–5934, 2005. DOI: 10.1158/1078-0432.CCR-05-0130.

[105] J. K. Aronson and A. R. Green, "Me-too pharmaceutical products: History, definitions, examples, and relevance to drug shortages and essential medicines lists," *British Journal of Clinical Pharmacology*, vol. 86, pp. 2114–2122, 2020. DOI: 10.1111/bcp.14327.

[106] M. Weber, M. Yurochkin, S. Botros, and V. Markov, "Black loans matter: Distributionally robust fairness for fighting subgroup discrimination," *CoRR*, vol. abs/2012.01193, 2020. arXiv: 2012.01193. [Online]. Available: https://arxiv.org/abs/2012.01193.

[107] H. Bandi and D. Bertsimas, "The price of diversity," *CoRR*, vol. abs/2107.03900, 2021. arXiv: 2107.03900. [Online]. Available: https://arxiv.org/abs/2107.03900.

[108] A. Lambrecht and C. Tucker, "Algorithmic bias? an empirical study of apparent gender-based discrimination in the display of stem career ads," *Management Science*, vol. 65, pp. 2947–3448, 2019. DOI: 10.1287/mnsc.2018.3093.

[109] M. A. Mazurowski, P. A. Habasa, J. M. Zurada, J. Y. Lo, J. A. Baker, and G. D. Tourassi, "Training neural network classifiers for medical decision making: The effects of imbalanced datasets on classification performance," *Neural Networks*, vol. 21, no. 2-3, pp. 2947–3448, 2008. DOI: 10.1016/j.neunet.2007.12.031.

[110] L. Seyyed-Kalantari, G. Liu, M. McDermott, I. Y. Chen, and M. Ghassemi, "Chexclusion: Fairness gaps in deep chest x-ray classifiers," *Biocomputing*, pp. 232–243, 2021. DOI: 10.1142/9789811232701_0022.

[111] Z. Obermeyer, B. Powers, C. Vogeli, and S. Mullainathan, "Dissecting racial bias in an algorithm used to manage the health of populations," *Science*, vol. 366, no. 6464, pp. 447–453, 2019. DOI: 10.1126/science.aax2342.

[112] R. A. Bauder, T. M. Khoshgoftaar, and T. Hasanin, "An empirical study on class rarity in big data," in *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)*, 2018, pp. 785–790. DOI: 10.1109/ICMLA.2018.00125.

[113] R. A. Bauder and T. M. Khoshgoftaar, "The effects of varying class distribution on learner behavior for medicare fraud detection with imbalanced big data," *Health Information Science and Systems*, vol. 6, no. 9, 2018. DOI: 10.1007/s13755-018-0051-3.

[114] Z.-H. Zhou and X.-Y. Liu, "Training cost-sensitive neural networks with methods addressing the class imbalance problem," *IEEE Transactions on Knowledge and Data Engineering*, vol. 18, no. 1, pp. 63–77, 2006. DOI: 10.1109/TKDE.2006.17.

[115] A. More, *Survey of resampling techniques for improving classification performance in unbalanced datasets*, 2016. DOI: 10.48550/ARXIV.1608.06048.

[116] P. Sattigeri, S. C. Hoffman, V. Chenthamarakshan, and K. R. Varshney, "Fairness gan: Generating datasets with fairness properties using a generative adversarial network," *IBM Journal of Research and Development*, vol. 63, no. 4/5, 3:1–3:9, 2019. DOI: 10.1147/JRD.2019.2945519.

[117] F. Calmon, D. Wei, B. Vinzamuri, K. Natesan Ramamurthy, and K. R. Varshney, "Optimized pre-processing for discrimination prevention," in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, *et al.*, Eds., vol. 30, Curran Associates, Inc., 2017.

[118] D. P. Kingma and M. Welling, *Auto-encoding variational bayes*, 2013. DOI: 10.48550/ARXIV.1312.6114. [Online]. Available: https://arxiv.org/abs/1312.6114.

[119] S. Kullback and R. Leibler, "On information and sufficiency," *The Annals of Mathematical Statistics*, vol. 22, no. 1, pp. 79–86, 1951. DOI: 10.1214/aoms/1177729694.

[120] T. M. Cover and P. E. Hart, "Nearest neighbor pattern classification," *IEEE Transactions on Information Theory*, vol. 13, no. 1, pp. 21–27, 1967. DOI: 10.1109/TIT.1967.1053964.

[121] J. Brabec and L. Machlica, "Bad practices in evaluation methodology relevant to class-imbalanced problems," *CoRR*, vol. abs/1812.01388, 2018. arXiv: 1812.01388. [Online]. Available: http://arxiv.org/abs/1812.01388.

[122] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep inside convolutional networks: Visualising image classification models and saliency maps," in *Workshop at International Conference on Learning Representations*, 2014. DOI: 10.48550/arXiv.1312.6034.

[123] L. van der Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of Machine Learning Research*, vol. 9, no. 86, 25792605, 2008.

[124] J. Krieger, D. Li, and D. Papanikolaou, "Missing Novelty in Drug Development*," *The Review of Financial Studies*, vol. 35, no. 2, pp. 636–679, Mar. 2021, ISSN: 0893-9454. DOI: 10.1093/rfs/hhab024. eprint: https://academic.oup.com/rfs/article-pdf/35/2/636/42228896/hhab024.pdf. [Online]. Available: https://doi.org/10.1093/rfs/hhab024.

[125] J. J. H. Park, E. Siden, M. J. Zoratti, *et al.*, "Systematic review of basket trials, umbrella trials, and platform trials: A landscape analysis of master protocols," *Trials*, vol. 20, no. 1, p. 572, 2019. DOI: 10.1186/s13063-019-3664-1.

[126] D. F. Heitjan, Z. Ge, and G.-s. Ying, "Real-time prediction of clinical trial enrollment and event counts: A review," *Contemporary Clinical Trials*, vol. 45, pp. 26–33, 2015. DOI: 10.1016/j.cct.2015.07.010.

[127] X. Zhang and Q. Long, "Modeling and prediction of subject accrual and event times in clinical trials: A systematic review," *Clinical Trials*, vol. 9, no. 6, pp. 681–688, 2012. DOI: 10.1177/1740774512447996.

[128] Y. J. Lee, "Interim recruitment goals in clinical trials," *Journal of Chronic Diseases*, vol. 36, pp. 379–389, 1983. DOI: 10.1016/0021-9681(83)90170-4.

[129] V. V. Anisimov and V. V. Fedorov, "Modelling, prediction and adaptive adjustment of recruitment in multicentre trials," *Statistics in Medicine*, vol. 26, pp. 4958–4975, 2007. DOI: 10.1002/sim.2956.

[130] J. Qian, D. K. Stangl, and S. George, "A weibull model for survival data: Using prediction to decide when to stop a clinical trial," in *Bayesian Biostatistics*, D. A. Berry and D. K. Stangl, Eds., Marcel Dekker, 1996, pp. 187–205.

[131] E. Bagiella and D. F. Heitjan, "Predicting analysis times in randomized clinical trials," *Statistics in Medicine*, vol. 40, pp. 2413–2421, 2001. DOI: 10.1002/sim.843.

[132] R. Machida, Y. Fujii, and T. Sozu, "Predicting study duration in clinical trials with a time-to-event endpoint," *Statistics in Medicine*, vol. 20, pp. 2055–2063, 2021. DOI: 10.1002/sim.8911.

[133] J. Liu, P. J. Allen, L. Benz, D. Blickstein, E. Okidi, and X. Shi, "A machine learning approach for recruitment prediction in clinical trial design," in *Proceedings of Machine Learning Research*, 2021. DOI: 10.48550/arXiv.2111.07407.

[134] T.-T. Chen, "A systematic review describes models for recruitment prediction at the design stage of a clinical trial," *BMC Medical Research Methodology*, vol. 16, p. 12, 2016. DOI: 10.1186/s12874-016-0117-3.

[135] E. Gkioni, R. Rius, S. Dodd, and C. Gamble, "A systematic review describes models for recruitment prediction at the design stage of a clinical trial," *Journal of Clinical Epidemiology*, vol. 115, pp. 141–149, 2019. DOI: 10.1016/j.jclinepi.2019.07.002.

[136] C. Davidson-Pilo, J. Kalderstam, P. Zivich, *et al.*, "Camdavidsonpilon/lifelines: V0.21.0," 2019. DOI: 10.5281/ZENODO.2638135.

[137] J. R. Quinlan, "Induction of decision trees," *Machine Learning*, vol. 1, pp. 81–106, 1986. DOI: 10.1007/BF00116251.

[138] S. Pölsterl, "Scikit-survival: A library for time-to-event analysis built on top of scikit-learn," *Journal of Machine Learning Research*, vol. 21, no. 212, pp. 1–6, 2020. DOI: 10.1186/s12874-016-0117-3.

[139] M. Leblanc and J. Crowley, "Survival trees by goodness of split," *Journal of the American Statistical Association*, vol. 88, no. 422, pp. 457–467, 1993. DOI: 10.1080/01621459.1993.10476296.

[140] H. Ishwaran, U. B. Kogalur, E. H. Blackstone, and M. S. Lauer, "Random survival forests," *Annals of Applied Statistics*, vol. 2, no. 3, pp. 841–860, 2008. DOI: 10.1214/08-AOAS169.

[141] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," *Annals of Statistics*, vol. 29, no. 5, pp. 1189–1232, 2001. DOI: 10.1214/aos/1013203451.

[142] T. Hothorn, P. Bühlmann, S. Dudoit, A. Molinaro, and M. J. Van Der Laan, "Survival ensembles," *Biostatistics*, vol. 7, no. 3, pp. 355–373, 2006. DOI: 10.1093/biostatistics/kxj011.

[143] Y. Goldberg, "Neural network methods for natural language processing," *Synthesis Lectures on Human Language Technologies*, vol. 10, pp. 1–309, 2017. DOI: 10.2200/S00762ED1V01Y201703HLT037.

[144] F. Sultana, A. Sufian, and P. Dutta, "Advancements in image classification using convolutional neural network," in *2018 Fourth International Conference on Research in Computational Intelligence and Communication Networks (ICRCICN)*, 2018, pp. 122–129. DOI: 10.1109/ICRCICN.2018.8718718.

[145] P. Bell, J. Fainberg, O. Klejch, J. Li, S. Renals, and P. Swietojanski, "Adaptation algorithms for neural network-based speech recognition: An overview," *IEEE Open Journal of Signal Processing*, vol. 2, pp. 33–66, 2021. DOI: 10.1109/OJSP.2020.3045349.

[146] J. L. Katzman, U. Shaham, A. Cloninger, J. Bates, T. Jiang, and Y. Kluger, "Deepsurv: Personalized treatment recommender system using a cox proportional hazards deep neural network," *BMC Medical Research Methodology*, vol. 18, p. 24, 2018. DOI: 10.1186/s12874-018-0482-1.

[147] A. F. Agarap, *Deep learning using rectified linear units (relu)*, 2018. DOI: 10.48550/ARXIV.1803.08375. [Online]. Available: https://arxiv.org/abs/1803.08375.

[148] D. P. Kingma and J. Ba, *Adam: A method for stochastic optimization*, 2014. DOI: 10.48550/ARXIV.1412.6980. [Online]. Available: https://arxiv.org/abs/1412.6980.

[149] S. Fotso, *Deep neural networks for survival analysis based on a multi-task framework*, 2018. DOI: 10.48550/ARXIV.1801.05512. [Online]. Available: https://arxiv.org/abs/1801.05512.

[150] S. Pölsterl, N. Navab, and A. Katouzian, "Fast training of support vector machines for survival analysis," in *Machine Learning and Knowledge Discovery in Databases*, A. Appice, P. P. Rodrigues, V. Santos Costa, J. Gama, A. Jorge, and C. Soares, Eds., Cham: Springer International Publishing, 2015, pp. 243–259, ISBN: 978-3-319-23525-7.

[151] A. R. Brentnall and J. Cuzick, "Use of the concordance index for predictors of censored survival data," *Statistical Methods in Medical Research*, vol. 27, no. 8, pp. 2359–2373, 2018. DOI: 10.1177/0962280216680245.

[152] S. Fotso *et al.*, *Pysurvival: Open source package for survival analysis modeling*, 2019. [Online]. Available: https://www.pysurvival.io/.

[153] Y. Saeys, T. Abeel, and Y. Van de Peer, "Robust feature selection using ensemble feature selection techniques," in *Machine Learning and Knowledge Discovery in Databases*, W. Daelemans, B. Goethals, and K. Morik, Eds., Berlin, Heidelberg: Springer Berlin Heidelberg, 2008, pp. 313–325, ISBN: 978-3-540-87481-2.

[154] A. Bashar, "Survey on evolving deep learning neural network architectures," *Journal of Artificial Intelligence and Capsule Networks*, vol. 1, no. 2, pp. 73–82, 2019. DOI: 10.36548/jaicn.2019.2.003.

[155] A. Altmann, L. Toloşi, O. Sander, and T. Lengauer, "Permutation importance: A corrected feature importance measure," *Bioinformatics*, vol. 26, no. 10, pp. 1340–1347, 2010. DOI: 10.1093/bioinformatics/btq134.

[156] P. L. Reiter, M. L. Katz, J. M. Oliveri, *et al.*, "Validation of self-reported colorectal cancer screening behaviors among appalachian residents," *Public Health Nurse*, vol. 30, no. 4, pp. 312–322, 2013. DOI: 10.1111/phn.12038.

[157] A. Sommer, J. Katz, and I. Tarwotjo, "Increased risk of respiratory disease and diarrhea in children with preexisting mild vitamin a deficiency," *The American journal of clinical nutrition*, vol. 40, no. 5, pp. 1090–1095, 1984. DOI: 10.1093/ajcn/40.5.1090.

[158] D. X. Li, "On default correlation: A copula function approach," *The Journal of Fixed Income*, vol. 9, pp. 43–54, 4 2000.

[159] R. Frey and A. J. McNeil, "Dependent defaults in models of portfolio credit risk," *Journal of Risk*, vol. 6, no. 1, pp. 59–92, 2003. DOI: 10.21314/JOR.2003.089.

[160] M. Noh, Y. Yip, Y. Lee, and Y. Pawitan, "Multicomponent variance estimation for binary traits in family-based studies," *Genetic Epidemiology*, vol. 30, no. 1, pp. 37–47, 2006.

[161] M. Noh and Y. Lee, "Reml estimation for binary data in glmms," *Journal of Multivariate Analysis*, vol. 98, pp. 896–915, 2007.

[162] M. C. Paik, "Repeated measurement analysis for nonnormal data in small samples," *Communications in Statistics – Simulation and Computation*, vol. 17, pp. 1155–1171, 1988.

[163] J. Yan, "Geepack: Yet another package for generalized estimating equations," *R-News*, vol. 2/3, pp. 12–14, 2002.

[164] U. Halekoh, S. Højsgaard, and J. Yan, "The r package geepack for generalized estimating equations," *Journal of Statistical Software*, vol. 15/2, pp. 1–11, 2006. DOI: 10.18637/jss.v015.i02.

[165] S. R. Collins, M. Z. Gunja, and M. M. Doty. "How well does insurance coverage protect consumers from health care costs?" (2017), [Online]. Available: https://www.commonwealthfund.org/publications/issue-briefs/2017/oct/

how-well-does-insurance-coverage-protect-consumers-health-care (visited on 02/14/2019).

[166] W. H. Shrank, N. K. Choudhry, M. A. Fischer, *et al.*, "The epidemiology of prescriptions abandoned at the pharmacy," *Annals of Internal Medicine*, vol. 153, no. 10, pp. 633–640, 2010. DOI: 10.7326/0003-4819-153-10-201011160-00005.

[167] U. D. of Health Human Services. "American patients first: The trump administration blueprint to lower drug prices and reduce out-of-pocket costs." (2018), [Online]. Available: https://www.hhs.gov/sites/default/files/AmericanPatientsFirst.pdf (visited on 02/14/2019).

[168] ——, "Fact sheet: Trump administration proposes to lower drug costs by targeting backdoor rebates and encouraging direct discounts to patients." (2019), [Online]. Available: https://www.hhs.gov/sites/default/files/20190131-fact-sheet.pdf (visited on 02/14/2019).

[169] N. C. Wright, A. C. Looker, K. G. Saag, *et al.*, "The recent prevalence of osteoporosis and low bone mass in the united states based on bone mineral density at the femoral neck or lumbar spine," *Journal of Bone and Mineral Research*, vol. 29, no. 11, pp. 2520–2526, 2014. DOI: 10.1002/jbmr.2269.

[170] A. Wilk, S. Sajjan, A. Modi, C. P. Fan, and P. Mavros, "Post-fracture pharmacotherapy for women with osteoporotic fracture: Analysis of a managed care population in the usa," *Osteoporosis International*, vol. 25, no. 12, pp. 2777–2786, 2014. DOI: 10.1007/s00198-014-2827-x.

[171] S. R. Cummings and L. J. Melton, "Epidemiology and outcomes of osteoporotic fractures," *Lancet*, vol. 358, no. 9319, pp. 1761–1767, 2002. DOI: 10.1016/S0140-6736(02)08657-9.

[172] D. M. Black and C. J. Rosen, "Postmenopausal osteoporosis," *New England Journal of Medicine*, vol. 374, no. 3, pp. 254–262, 2016. DOI: 10.1056/NEJMcp1513724.

[173] M. Derbyshire, "Patent expiry dates for biologicals: 2017 update," *Generics and Biosimilars Initiative Journal*, vol. 7, no. 1, pp. 29–34, 2018. DOI: 10.5639/gabij.2018.0701.007.

[174] G. Dagotto, J. Yu, and D. H. Barouch, "Approaches and challenges in sars-cov-2 vaccine development," *Cell Host Microbe*, vol. 28, no. 3, pp. 364–370, 2020. DOI: 10.1016/j.chom.2020.08.002.

[175] S. K. Shah, F. G. Miller, T. C. Darton, *et al.*, "Ethics of controlled human infection to address covid-19," *Science*, vol. 368, no. 6493, pp. 832–834, 2020. DOI: 10.1126/science.abc1076.

[176] S. A. Plotkin and A. Caplan, "Extraordinary diseases require extraordinary solutions," *Vaccine*, vol. 38, no. 24, pp. 3987–3988, 2020. DOI: 10.1016/j.vaccine.2020.04.039.

[177] N. Eyal, M. Lipsitch, and P. G. Smith, "Human challenge studies to accelerate coronavirus vaccine licensure," *The Journal of Infectious Diseases*, vol. 221, no. 11, pp. 1752–1756, 2020. DOI: 10.1093/infdis/jiaa152.

[178] C. P. Gross and K. A. Sepkowitz, "The myth of the medical breakthrough: Smallpox, vaccination, and jenner reconsidered," *The Journal of Infectious Diseases*, vol. 3, no. 1, pp. 54–60, 2020. DOI: 10.1093/infdis/jiaa152.

[179] E. Jamrozik and M. J. Selgelid, "History of human challenge studies," in *Human Challenge Studies in Endemic Settings: Ethical and Regulatory Issues*. Cham: Springer International Publishing, 2021, pp. 9–23, ISBN: 978-3-030-41480-1. DOI: 10.1007/978-3-030-41480-1_2".

[180] S. Riedel, "Edward jenner and the history of smallpox and vaccination," *Proceedings (Baylor University. Medical Center)*, vol. 18, no. 1, pp. 21–25, 2005. DOI: 10.1080/08998280.2005.11928028.

[181] D. P. Steensma, V. M. Montori, M. A. Shampo, and R. A. Kyle, "Daniel alcides carrión – peruvian hero and medical martyr," *Proceedings (Baylor University.*

Medical Center), vol. 89, no. 6, e55–e56, 2014. DOI: 10.1016/j.mayocp.2013.08.025.

[182] A. Mehra, "Politics of participation: Walter reed's yellow-fever experiments," *Virtual Mentor*, vol. 11, no. 4, pp. 326–330, 2009. DOI: 10.1001/virtualmentor.2009.11.4.mhst1-0904.

[183] A. A. Smorodintseff, M. D. Tushinsky, A. I. Drobyshevskaya, A. A. Korovin, and A. I. Osetroff, "Investigation on volunteers infected with the influenza virus," *American Journal of Medical Sciences*, vol. 194, pp. 159–170, 1937.

[184] G. Baader, S. E. Lederer, M. Low, F. Schmaltz, and A. V. Schwerin, "Pathways to human experimentation," *Osiris*, vol. 20, pp. 205–231, 2005. DOI: 10.1086/649419.

[185] S. M. Reverby, "Ethical failures and history lessons: The u.s. public health service research studies in tuskegee and guatemala," *Public Health Reviews*, vol. 34, p. 13, 2012. DOI: 10.1007/BF03391665.

[186] M. L. Clements, R. F. Betts, and B. R. Murphy, "Advantage of live attenuated cold-adapted influenza a virus over inactivated vaccine for a/washington/80 (h3n2) wild-type virus infection," *Lancet*, vol. 1, no. 8379, pp. 705–708, 1984. DOI: 10.1016/s0140-6736(84)92222-0.

[187] J. F. Mosley 2nd, L. L. Smith, P. Brantley, D. Locke, and M. Como, "The first fda-approved cholera vaccination in the united states," *Pharmacy and Therapeutics*, vol. 42, no. 10, pp. 638–640, 2017.

[188] C. Jin, M. M. Gibani, M. Moore, *et al.*, "Efficacy and immunogenicity of a vi-tetanus toxoid conjugate vaccine in the prevention of typhoid fever using a controlled human infection model of salmonella typhi: A randomised controlled, phase 2b trial," *Lancet*, vol. 390, no. 10111, pp. 2472–2480, 2017. DOI: 0.1016/S0140-6736(17)32149-9.

[189] S. Mahmoudi and H. Keshavarz, "Efficacy of phase 3 trial of rts, s/as01 malaria vaccine: The need for an alternative development plan," *Human Vaccines Immunotherapeutics*, vol. 13, no. 9, pp. 2098–2101, 2017. DOI: 10.1080/21645515.2017.1295906.

[190] R. Meta, M.-A. Hoogerwerf, D. M. Ferreira, B. Mordmüller, and M. Yazdanbakhsh, "Experimental infection of human volunteers," *Lancet Infectious Diseases*, vol. 18, no. 10, e321–e322, 2018. DOI: 10.1016/S1473-3099(18)30177-4.

[191] A. C. Sherman, A. Mehta, N. W. Dickert, E. J. Anderson, and N. 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*, vol. 9, p. 107, 2019. DOI: 10.3389/fcimb.2019.00107.

[192] M. G. Ison, V. Campbell, C. Rembold, J. Dent, and F. G. Hayden, "Cardiac findings during uncomplicated acute influenza in ambulatory adults," *Clinical Infectious Diseases*, vol. 40, no. 3, pp. 415–422, 2005. DOI: 10.1086/427282.

[193] M. J. Memoli, L. Czajkowski, S. Reed, *et al.*, "Validation of the wild-type influenza a human challenge model h1n1pdmist: An a(h1n1)pdm09 dose-finding investigational new drug study," *Clinical Infectious Diseases*, vol. 60, no. 5, pp. 693–702, 2015. DOI: 10.1093/cid/ciu924.

[194] Council for International Organizations of Medical Sciences. "International Ethical Guidelines for Health-related Research Involving Humans - 2016." (2017), [Online]. Available: https://cioms.ch/wp-content/uploads/2017/01/WEB-CIOMS-EthicalGuidelines.pdf (visited on 11/18/2021).

[195] U.S. National Institute of Allergy and Infectious Diseases. "Ethical Considerations for Zika Virus Human Challenge Trials." (2017), [Online]. Available: https://www.niaid.nih.gov/sites/default/files/EthicsZikaHumanChallengeStudiesRe
pdf (visited on 02/01/2017).

[196] R. Palacios and S. K. Shah, "When could human challenge trials be deployed to combat emerging infectious diseases? lessons from the case of a zika virus human challenge trial," *Trials*, vol. 20, p. 702, 2019. DOI: 10.1186/s13063-019-3843-0.

[197] J. P. Kahn, L. M. Henry, A. C. Mastroianni, W. H. Chen, and R. Macklin, "For now, it's unethical to use human challenge studies for sars-cov-2 vaccine development," *Clinical Infectious Diseases*, vol. 117, no. 46, pp. 28 538–28 542, 2020. DOI: 10.1073/pnas.2021189117.

[198] M. E. Deming, N. L. Michael, M. Robb, M. S. Cohen, and K. M. Neuzil, "Accelerating development of sars-cov-2 vaccines – the role for controlled human infection models," *New England Journal of Medicine*, vol. 383, no. 10, e63, 2020. DOI: 10.1056/NEJMp2020076.

[199] S. Holm, "Controlled human infection with sars-cov-2 to study covid-19 vaccines and treatments: Bioethics in utopia," *Journal of Medical Ethics*, vol. 46, no. 9, pp. 569–573, 2020. DOI: 10.1136/medethics-2020-106476.

[200] E. Jamrozik and M. J. Selgelid, "Covid-19 human challenge studies: Ethical issues," *Lancet Infectious Diseases*, vol. 20, no. 8, e198–e203, 2020. DOI: 10.1016/S1473-3099(20)30438-2.

[201] L. Dawson, J. Earl, and J. Livezey, "Severe acute respiratory syndrome coronavirus 2 human challenge trials: Too risky, too soon," *Journal of Infectious Diseases*, vol. 222, no. 3, pp. 514–516, 2020. DOI: 10.1093/infdis/jiaa314.

[202] D. P. Sulmasy, "Are sars-cov-2 human challenge trials ethical?" *JAMA Internal Medicine*, vol. 181, no. 8, pp. 1031–1032, 2021. DOI: 10.1001/jamainternmed.2021.2614.

[203] A. M. Capron, "Looking back at withdrawal of life-support law and policy to see what lies ahead for medical aid-in-dying," *Yale Journal of Biology and Medicine*, vol. 92, no. 4, pp. 781–791, 2019.

[204] E. Callaway, "Dozens to be deliberately infected with coronavirus in uk 'human challenge' trials," *Nature*, vol. 586, pp. 651–652, 2020. DOI: 10.1038/d41586-020-02821-4.

[205] G. Rapeport, E. Smith, A. Gilbert, A. Catchpole, M. F. Helen, and C. Chiu, "Sars-cov-2 human challenge studies – establishing the model during an evolving pandemic," *New England Journal of Medicine*, vol. 385, no. 11, pp. 961–964, 2021. DOI: 10.1056/NEJMp2106970.

[206] D. A. Berry, "Adaptive clinical trials in oncology," *Nature Review Clinical Oncology*, vol. 9, no. 4, pp. 199–207, 2012. DOI: 10.1038/nrclinonc.2011.165.

[207] S. M. Paul, D. S. Mytelka, C. T. Dunwiddie, *et al.*, "How to improve rd productivity: The pharmaceutical industry's grand challenge," *Nature Reviews Drug Discovery*, vol. 9, no. 3, pp. 203–214, 2010. DOI: 10.1038/nrd3078.

[208] J. Strovel, S. Sittampalam, N. P. Coussens, *et al.* "Early drug discovery and development guidelines: For academic researchers, collaborators, and start-up companies, 2016." (2016), [Online]. Available: https://www.ncbi.nlm.nih.gov/books/NBK92015/ (visited on 07/01/2016).

[209] K. I. Kaitin and J. A. DiMasi, "Pharmaceutical innovation in the 21st century: New drug approvals in the first decade, 2000–2009," *Clinical Pharmacology & Therapeutics*, vol. 89, no. 2, pp. 183–188, 2011. DOI: 10.1038/clpt.2010.286.

[210] B. Research. "Brain tumor therapeutics: Global markets to 2023." (2019), [Online]. Available: https://www.bccresearch.com/market-research/pharmaceuticals/brain-tumor-therapeutics-markets.html (visited on 05/03/2019).

[211] A. H. I. P. C. for Policy and Research. "High-priced drugs: Estimates of annual per-patient expenditures for 150 specialty medications." (2016), [Online]. Available: https://www.ahip.org/wp-content/uploads/2016/04/HighPriceDrugsReport.pdf (visited on 04/04/2016).