Statistical Foundations for Precision Medicine

by

Arjun Kumar Manrai

A.B., Physics
Harvard College (2008)

Submitted to the Harvard-MIT Division of Health Sciences and Technology
In Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

FEBRUARY 2015

Signature redacted

Signature of Author...
Harvard-MIT Division of Health Sciences and Technology
February 2015

Signature redacted

Certified by...
Isaac S. Kohane, M.D., Ph.D.
Professor of Pediatrics and Health Sciences and Technology
Thesis Supervisor

Accepted by.... Signature redacted ....
Emery N. Brown, M.D., Ph.D.
Director, Harvard-MIT Program in Health Sciences and Technology
Professor of Computational Neuroscience and Health Sciences and Technology

This page is intentionally left blank.

Statistical Foundations for Precision Medicine

by

Arjun Kumar Manrai

Submitted to the Harvard-MIT Division of Health Sciences and Technology on February 2, 2015
In Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy

Abstract

Physicians must often diagnose their patients using disease archetypes that are based on symptoms as opposed to underlying pathophysiology. The growing concept of "precision medicine" addresses this challenge by recognizing the vast yet fractured state of biomedical data, and calls for a patient-centered view of data in which molecular, clinical, and environmental measurements are stored in large shareable databases. Such efforts have already enabled large-scale knowledge advancement, but they also risk enabling large-scale misuse. In this thesis, I explore several statistical opportunities and challenges central to clinical decision-making and knowledge advancement with these resources. I use the inherited heart disease hypertrophic cardiomyopathy (HCM) to illustrate these concepts.

HCM has proven tractable to genomic sequencing, which guides risk stratification for family members and tailors therapy for some patients. However, these benefits carry risks. I show how genomic misclassifications can disproportionately affect African Americans, amplifying healthcare disparities. These findings highlight the value of diverse population sequencing data, which can prevent variant misclassifications by identifying ancestry informative yet clinically uninformative markers. As decision-making for the individual patient follows from knowledge discovery by the community, I introduce a new quantity called the "dataset positive predictive value" (dPPV) to quantify reproducibility when many research teams separately mine a shared dataset, a growing practice that mirrors genomic testing in scale but not synchrony. I address only a few of the many challenges of delivering sound interpretation of genetic variation in the clinic and the challenges of knowledge discovery with shared "big data." These examples nonetheless serve to illustrate the need for grounded statistical approaches to reliably use these powerful new resources.

Thesis Supervisor: Isaac S. Kohane, M.D., Ph.D.
Title: Professor of Pediatrics and Health Sciences and Technology

Acknowledgments

This work would have been impossible without the support of many people and it is a privilege to have been in their company for the past several years. I owe my deepest gratitude to the following individuals:

Dr. Isaac (Zak) Kohane, my brilliant mentor and friend. When I first met Zak, I knew I needed to work for him. I left our first meeting (and every meeting thereafter) feeling inspired by his vision for medicine, and supported by his mentoring style which allowed me to ask and answer questions with complete freedom. I always felt that Zak's primary goal was to cultivate my research skills and hone my interests (and to make me laugh). He provided me with far more intellectual freedom than I deserved during graduate school, yet also held my work to highest standards. It is my good fortune to have spent so much time with such a nurturing adviser and prophetic thinker.

Dr. Peter Szolovits and Dr. Joseph Loscalzo, my committee, for their endless support, challenging questions, and willingness to always share their wisdom. They both have had a profound impact on the way I approach science.

My colleagues in ZakLab for many spirited and productive discussions over the years.

My collaborators outside ZakLab: the Lab for Molecular Medicine, Dr. Judith Strymish, and Dr. Sachin Jain—this work would not have been possible without all of you. A special thanks goes to Dr. John Ioannidis, one of my personal heroes and the inspiration for Chapters 3 and 4.

My friends during graduate school, especially Ally, Kartik, Adeeti, Gaurav, and Chirag—you all are the best.

My Rachna, for being my constant source of happiness.

My sister Mini, for always believing in me and having my back.

And finally, I dedicate this thesis to my parents, Drs. Lalita and Ajay Manrai, who have given me everything, even my curiosity and persistence. They both completed their PhDs while raising two young kids in two cities in a new country. I hope one day to achieve half of what they have both as parents and as scholars.

Table of Contents

This page is intentionally left blank.

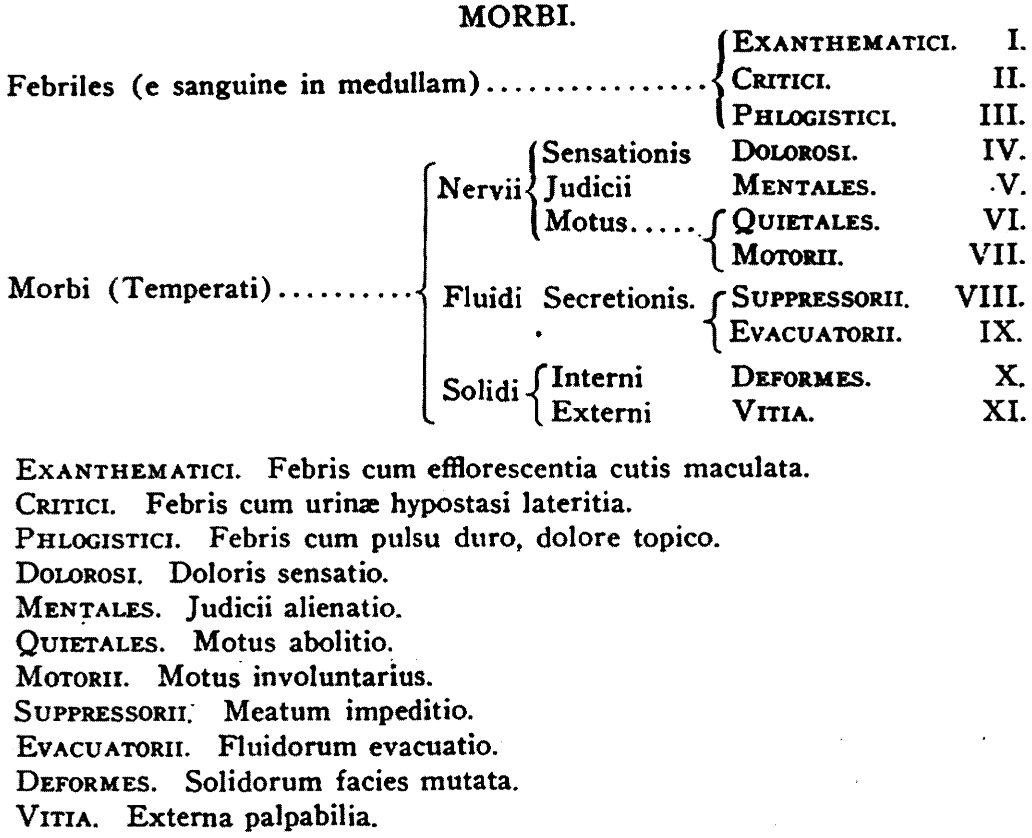
Chapter 1: Introduction

Medicine is a science of uncertainty and an art of probability.

-Sir William Osler

1.1 Precision Medicine

In 1763, Carl Linnaeus, the father of modern taxonomy, developed a classification system for human disease—a "nosology"—in his *Genera Morborum*.[1] Linnaeus' nosology contained 11 classes, 37 orders, and 325 "species" of disease. Nine of the eleven classes were based on symptoms, while the two classes *Deformes* and *Vitia* were based on anatomic findings (Figure 1). Linnaeus' classification system was a major contribution to medicine, especially in view of the understanding of disease etiology and pathophysiology during his time.[2]



MORBI.

Febriles (e sanguine in medullam)			Exanthematici.	I.
			Critici.	II.
			Phlogistici.	III.
Morbi (Temperati)	Nervii	Sensationis	Dolorosi.	IV.
		Judicii	Mentales.	V.
		Motus	Quietales.	VI.
			Motorii.	VII.
	Fluidi	Secretionis.	Suppressorii.	VIII.
			Evacuatorii.	IX.
	Solidi	Interni	Deformes.	X.
		Externi	Vitia.	XI.

Exanthematici. Febris cum efflorescentia cutis maculata.
Critici. Febris cum urinæ hypostasi lateritia.
Phlogistici. Febris cum pulsu duro, dolore topico.
Dolorosi. Doloris sensatio.
Mentales. Judicii alienatio.
Quietales. Motus abolitio.
Motorii. Motus involuntarius.
Suppressorii. Meatum impeditio.
Evacuatorii. Fluidorum evacuatio.
Deformes. Solidorum facies mutata.
Vitia. Externa palpabilia.

Figure 1: Carl Linnaeus' classification system for human disease published in *Genera Morborum*, 1763. Eleven classes of disease are listed on the right hand side of the figure (Roman numerals). Figure from Egdhal.[2]

Today, the most widely used taxonomy of human disease is the International Classification of Disease (ICD), a nosology considerably more complex than Linnaeus' version and an integral part of our healthcare system, used in myriad purposes from medical billing to research.[3] Epidemiologist Robert Hahn explains that the ICD is intended "to include all conditions—and to ensure that no particular event of sickness will be classified under more than one code number."[4] To accomplish these goals, the ICD includes clauses that explicitly exclude conditions (e.g. M76: "Enthesopathies, lower limb, excluding foot") and others that leave room for future refinement (F84.9: "Pervasive developmental disorder, unspecified"). Yet while this nosology is more than 250 years removed from Linnaeus and we are more than a decade into the genomics era, the diseases in the ICD themselves—Linnaeus' "species"—are still largely diagnosed by symptoms, signs, and simple Oslerian clinicopathological correlations[5] as opposed to the patient's underlying and more complex pathophysiology.[6]

Consider the classification of type 2 diabetes. This disease is usually diagnosed by an abnormal fasting blood glucose or by an abnormal three-month average of blood glucose, HbA1c.[7] This downstream physiological response (impaired glucose tolerance) may be treated pharmacologically with oral medication (e.g. Glucophage) or intramuscularly with insulin.[8] Such strategies have markedly improved the quality and longevity of life for those with type 2 diabetes. Notwithstanding these accomplishments, we know this disease has a multifactorial etiology with a strong hereditary component,[9] but we have little understanding of its underlying molecular pathophysiology and the reasons that the disease manifests so heterogeneously across individuals. Thus, we are left to manage the symptoms and their downstream consequences (e.g. retinopathy[10], nephropathy[11]) without the ability to detect or treat its upstream cause and without a precise description of disease progression or risk for family members. Decades of clinical experience have shown that it is crucial for diabetes, and even for many ostensibly "single gene" disorders, to contextualize the disease using its multiple genomic and environmental determinants, as depicted in Figure 2.[12-15]

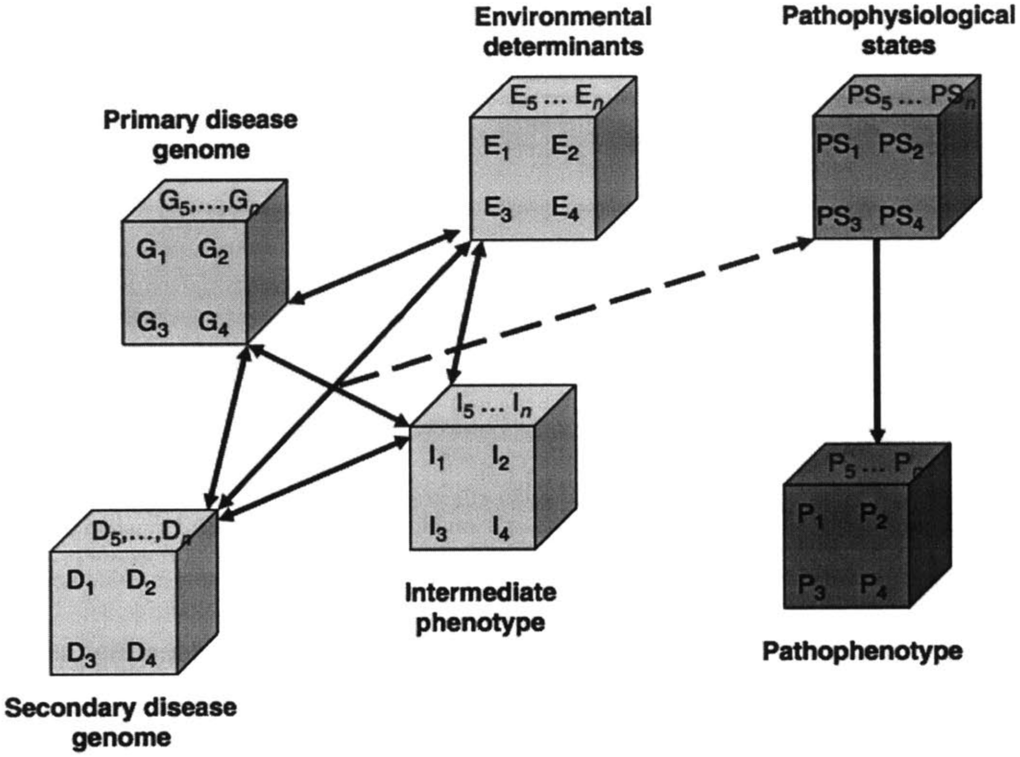


Figure 2: Human disease network. The primary disease genome, secondary disease genome, environmental determinants, and intermediate phenotype interact to yield pathophysiological states and pathophenotypes. Figure from Loscalzo et al.[3]

The concept of "precision medicine" formalizes the challenge of reclassifying disease in the context of large-scale molecular and patient data generated by contemporary healthcare and biomedical research enterprises. This vision for medicine was described in a report published by the Committee on a Framework for

Developing a New Taxonomy of Disease ("Committee"), and calls for a taxonomy of disease that views a patient's state as a time-varying high-dimensional vector of genetic, environmental, and clinical data.[6] In order to achieve this new taxonomy, the Committee charges the research community to integrate the vast yet fragmented patient-centric resources into an "Information Commons" as depicted in Figure 3.

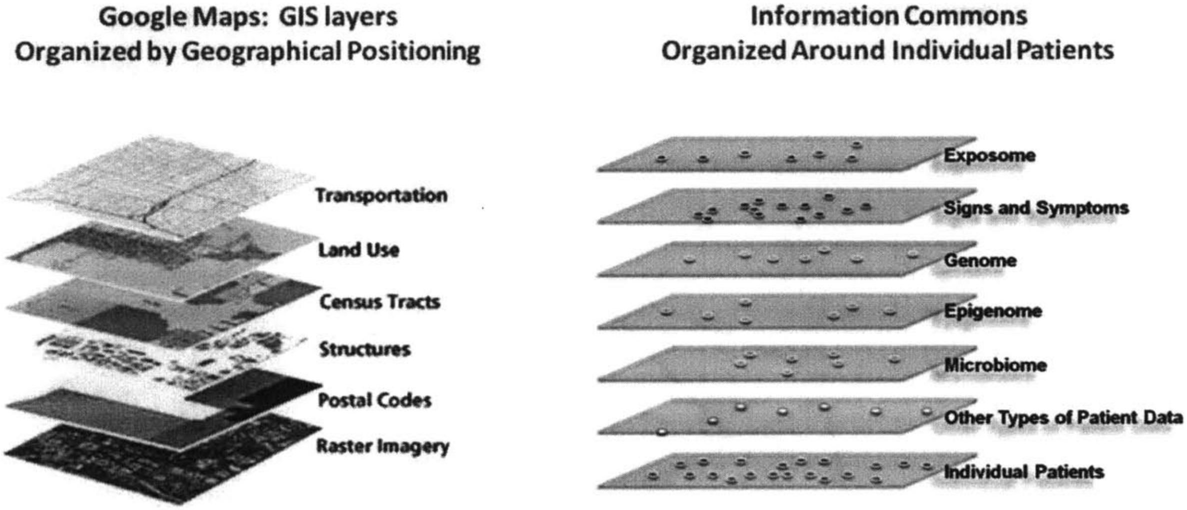


Figure 3: A patient-centric information commons is the data substrate for precision medicine. Just as Geographical Information Systems (GIS) are location-centric integrations of multiple layers of data, the Information Commons is a patient-centric integration of diverse data that collectively inform patient state. Figure from *Toward Precision Medicine* report.[6]

The Information Commons (right panel of Figure 3) bears resemblance to Geographical Information Systems (GIS, left panel of Figure 3). Just as GIS are physical location-centric integrations of multiple layers of data, the Information Commons is a patient-centric integration of diverse data types that collectively represent the patient's state. Sharing and centralizing these data will enhance efforts to reclassify ostensibly singular common phenotypes into their distinct constituent diseases, an approach that has already led to significant gains in the clinical management of diseases such as non-small cell lung carcinoma.[16]

1.2 Statistical Foundations for Precision Medicine

In this thesis, I explore several statistical opportunities and challenges central to clinical decision-making and knowledge advancement with growing "big data" resources. I focus on two ongoing use cases: (1) clinical interpretation of genetic variation and (2) knowledge advancement in the context of many researchers mining a shared dataset. A primary goal of "precision medicine" is to identify the true (and often distinct) *causes* of an apparently singular phenotype, but it is necessary to distinguish this long-term goal from ongoing clinical decision-making, which often requires only an accurate understanding of the *correlation* between genotype, phenotype, and other available data. If "precision" in describing such correlative evidence is not achieved, then the consequences for patients can be harmful, and public and private investments in biomedical research wasteful.

The threat of large-scale misuse in the context of genomic medicine was described nearly a decade ago by Kohane and colleagues, who defined "the incidentalome" in 2006 as the set of incidental findings obtained from comprehensive genotyping in the general population which, if unchecked, may largely be composed of false positives (Figure 4).[17]

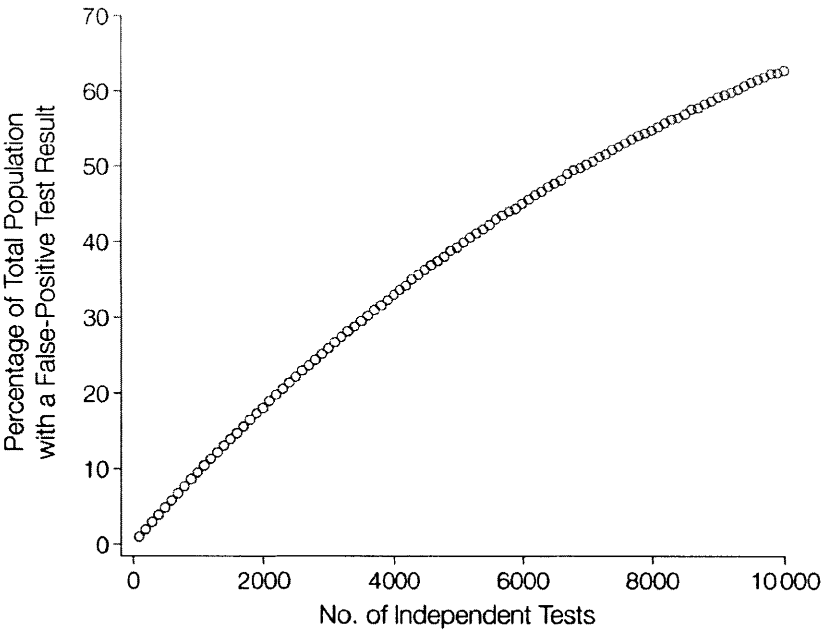


Figure 4: The percentage of the total population with a false-positive test result when genomic testing is applied to the general population for a large number of low-probability conditions. Figure from Kohane et al.[17]

Kohane and colleagues warn that the accumulation of false positive test results in the general population may be disastrously high even when genomic tests have nominally very good sensitivity and specificity if the tested population has low prior probability for the conditions tested. The rate at which false positive test results grow with respect to the number of independent tests depends on the clinical specificity (and sensitivity) of the tests—if the tests had perfect specificity, genomic variants would be pathognomonic ("patho*genomic*") with disease and false positives would be nonexistent. But much recent experience has shown that high specificity is the exception—reduced penetrance and variable expressivity are the norm.[18-20] This begs the question: what is the typical false positive rate? We currently lack the infrastructure to answer this question systematically across diseases but the Information Commons is poised to rapidly accelerate this goal. In order to enhance these future efforts, we identify three statistical challenges for ongoing clinical decision-making using genomic data as well as knowledge advancement using shared "big data" resources:

1. The bias-variance tradeoff is fundamental to precision medicine. On the one hand, if our understanding of disease is too coarse, we may fail to stratify diagnosis along meaningful axes (e.g. subclinical findings or ethnicity, where allele frequency differences between populations can masquerade as meaningful clinical signal[21]). On the other hand, it is remarkably easy in today's rich data environment to unwittingly use high-dimensional patient data to support virtually any hypothesis with biological plausibility. Finding the right balance, as with any statistical model, will lead to the most reliable findings outside the training data.

2. Communal science – Shared big data resources permit a multiplicity of uncoordinated investigations. This multiplicity rivals high-throughput omics multiple hypothesis-testing in scale but not synchrony: whereas high-throughput omics analyses often use a single test platform where all measurements are taken simultaneously, analyses using shared data may happen piece-meal by investigators over decades. New knowledge will need to be contextualized accordingly.

3. Medical education – Even with accurate estimates of the relationships between features across the layers of the Information Commons, new diagnostics and therapeutics will improve care only if their performance parameters (sensitivity, specificity) are readily available to physicians at the point of care and incorporated correctly into decision making. Statistical

literacy will become increasingly important if we are to keep pace with an ever-growing catalogue of diagnostics and therapeutics.[21]

The three challenges above are central to sound clinical interpretation of genomic data and knowledge advancement using shared big data resources.

1.3 Outline of Thesis

In Chapter 2, I describe a cautionary tale for genomic medicine that illustrates the importance of diverse control sequence data in the clinical interpretation of genetic variation for the inherited cardiac disease hypertrophic cardiomyopathy (HCM). In Chapters 3 and 4, I introduce a new quantity called the "dataset positive predictive value" (*dPPV*) to quantify the proportion of true claims amongst all claims made during multiple uses of a shared dataset by different investigators. In Chapter 5, I report the results of a study that assessed statistical literacy in a group of practicing physicians. In Chapter 6, I describe future directions of this research program.

Chapter 2 is based on a manuscript co-authored with B.H. Funke, H.L. Rehm, M.S. Olesen, B.A. Maron, P. Szolovits, D.M. Margulies, J. Loscalzo, and I.S. Kohane. Chapters 3 and 4 are based on a manuscript co-authored with C.J. Patel, J.P.A. Ioannidis, and I.S. Kohane. Chapter 5 is based on a manuscript co-authored with G. Bhatia, J. Strymish, I.S. Kohane, and S.H. Jain published in *JAMA Internal Medicine*.[22]

REFERENCES

1. Linnaeus, C. Genera Morborum. (1763).
2. Egdahl, A. Linnæus' 'Genera Morborum,' and Some of His Other Medical Works. *Medical Library and Historical Journal* **5**, 185 (1907).
3. Kohane, I.S. Using electronic health records to drive discovery in disease genomics. *Nat. Rev. Genet.* **12**, 417–428 (2011).
4. Hahn, R.A. *Sickness and Healing*. 336 (Yale University Press: 1996).
5. Loscalzo, J. & Barabasi, A.-L. Systems biology and the future of medicine. **3**, (2011).
6. *Toward Precision Medicine: Building a Knowledge Network for Biomedical Research and a New Taxonomy of Disease*. 121 (National Academies Press: 2011).
7. Alberti, K.G. & Zimmet, P.Z. Definition, diagnosis and classification of diabetes mellitus and its complications. Part 1: diagnosis and classification of diabetes mellitus provisional report of a WHO consultation. *Diabet. Med.* **15**, 539–553 (1998).
8. Nathan, D.M. et al. Medical management of hyperglycemia in type 2 diabetes: a consensus algorithm for the initiation and adjustment of therapy: a consensus statement of the American Diabetes Association and the European Association for the Study of Diabetes. *Diabetes Care* **32**, 193–203 (2009).
9. Poulsen, P., Kyvik, K.O., Vaag, A. & Beck-Nielsen, H. Heritability of type II (non-insulin-dependent) diabetes mellitus and abnormal glucose tolerance--a population-based twin study. *Diabetologia* **42**, 139–145 (1999).

10. Mohamed, Q., Gillies, M.C. & Wong, T.Y. Management of diabetic retinopathy: a systematic review. *JAMA* **298**, 902–916 (2007).

11. Mogensen, C.E. & Christensen, C.K. Predicting diabetic nephropathy in insulin-dependent patients. *New England Journal of Medicine* **311**, 89–93 (1984).

12. Loscalzo, J., Kohane, I. & Barabasi, A.-L. Human disease classification in the postgenomic era: a complex systems approach to human pathobiology. *Mol. Syst. Biol.* **3**, 124 (2007).

13. Barabasi, A.-L., Gulbahce, N. & Loscalzo, J. Network medicine: a network-based approach to human disease. *Nat. Rev. Genet.* **12**, 56–68 (2011).

14. Kato, G.J., Gladwin, M.T. & Steinberg, M.H. Deconstructing sickle cell disease: reappraisal of the role of hemolysis in the development of clinical subphenotypes. *Blood Rev.* **21**, 37–47 (2007).

15. Farber, H.W. & Loscalzo, J. Pulmonary arterial hypertension. *N. Engl. J. Med.* **351**, 1655–1665 (2004).

16. Pao, W. & Girard, N. New driver mutations in non-small-cell lung cancer. *Lancet Oncol.* **12**, 175–180 (2011).

17. Kohane, I.S., Masys, D.R. & Altman, R.B. The incidentalome: a threat to genomic medicine. *JAMA* **296**, 212–215 (2006).

18. Andreasen, C. et al. New population-based exome data are questioning the pathogenicity of previously cardiomyopathy-associated genetic variants. *Eur J Hum Genet* **21**, 918–928 (2013).

19. Lopes, L.R., Rahman, M.S. & Elliott, P.M. A systematic review and meta-analysis of genotype-phenotype associations in patients with hypertrophic

cardiomyopathy caused by sarcomeric protein mutations. *Heart* **99**, 1800–1811 (2013).

20. Rehm, H.L. Disease-targeted sequencing: a cornerstone in the clinic. *Nat. Rev. Genet.* **14**, 295–300 (2013).

21. Rubinstein, W.S. et al. The NIH genetic testing registry: a new, centralized database of genetic tests to enable access to comprehensive information and improve transparency. *Nucleic Acids Research* **41**, D925–D935 (2012).

22. Manrai, A.K., Bhatia, G., Strymish, J., Kohane, I.S. & Jain, S.H. Medicine's uncomfortable relationship with math: calculating positive predictive value. *JAMA Intern Med* **174**, 991–993 (2014).

This page is intentionally left blank.

Chapter 2

A Cautionary Tale for Genomic Medicine: Population Diversity and the Genetics of Hypertrophic Cardiomyopathy

Overview

Risk stratification for hypertrophic cardiomyopathy (HCM) is an exemplar of the clinical gains attainable by targeted genetic testing. Using sequencing results, clinicians routinely assess risk for the patient's relatives and even tailor therapy for rare patients. However, the benefits of genetic testing come with the risk that variants may be misclassified. Using publicly accessible exome data, we identified variants previously considered causal of HCM that were overrepresented in the general population. We studied these variants in diverse populations, and reevaluated their initial ascertainments in the medical literature. We reviewed patient records at a leading genetic testing laboratory for variant occurrences during the near decade-long history of the laboratory. Multiple patients, all of African or unspecified ancestry, received positive reports with variants initially classified as pathogenic and later changed to benign. All studied high-frequency variants were significantly more common in African Americans than European Americans ($P < 0.001$). If diverse control sequencing data had been available, these variants would likely have been classified earlier as benign, possibly avoiding multiple misclassifications in African-ancestry individuals. We identify methodological shortcomings that may have led to these errors in the medical literature. These findings highlight the value of diverse population sequencing data, which can prevent variant misclassifications by identifying ancestry informative yet clinically uninformative markers. These findings expand upon current guidelines, which recommend using ethnically matched controls to interpret variants. As diverse sequencing data become more widely available, we expect variant reclassifications to increase, particularly for ancestry groups that have historically been less well studied.

INTRODUCTION

Although hypertrophic cardiomyopathy (HCM) is best known as a fatal affliction of young athletes, it causes significant morbidity and mortality in patients of all ages and lifestyles.[1,2] The defining feature of HCM is unexplained left ventricular hypertrophy (LVH) but its clinical presentation is heterogeneous, manifesting as severe heart failure in some patients yet being asymptomatic in others.[3] In over one-third of patients, causal genetic lesions are identified, enabling clinicians to risk stratify the patient's relatives[4] and in specific, rare circumstances, tailor therapy for a patient found to have a tractable phenocopy disorder such as Fabry disease.[5] Additionally, in patients with clinical features but not a definitive diagnosis of HCM, identification of a pathogenic sarcomeric variant may be used to help establish a diagnosis.

When a patient is incorrectly informed that one of his or her variants is causal when in fact it is benign, it can have far-reaching unintended consequences within the family. First, relatives who lack the non-causal variant are given false reassurance that further surveillance is unnecessary. Second, relatives possessing the non-causal variant receive prolonged at-risk screening and are advised on lifestyle modifications (e.g., cessation of certain sports and activities) that may not be necessary, in addition to the stress and economic burden that accompany the incorrect diagnosis. Third, for patients with clinical features but without a definitive diagnosis of HCM, such as young athletes with modest hypertrophy and a family history of sudden cardiac death, misclassification of a benign sarcomeric variant as pathogenic may lead to overestimation of the benefits of implanting a cardioverter defibrillator to prevent sudden cardiac death. Lastly, when a variant's status is downgraded from pathogenic to benign, the sequencing laboratory often re-contacts the referring physician who, in turn, re-contacts the patient and their tested family members, engendering confusion and compromising trust.

In order to safeguard against the many problems that result from variant misclassifications, much effort has gone into developing standards for correct interpretation.[1,4,6-9] The principal challenge is to separate truly pathogenic variants from the historically underappreciated amount of background variant noise dormant in the genome.[6,10] To aid with interpretation, expert guidelines generally recommend classifying variants using ethnically matched control sequence data.[4,8].

Recently, large-scale control sequence data from the NHLBI Exome Sequence Project[11] were systematically reviewed for HCM-associated variants labeled "disease-causing" or "pathogenic" in an expert-curated database.[12,13] Far more HCM variants were found than expected in the general population, implying reduced penetrance or misclassification errors in prior HCM-variant associations, or both. We observed that only a handful of high-frequency variants account for the majority of this overabundance, and that these variants occur disproportionately in African American individuals.

We hypothesized that the identification of HCM-associated high-frequency variants in the general population implied historical reporting errors in patients, and that most or all individuals affected would be of African ancestry. We further posited that these variant associations stemmed from ascertainment bias and other methodological shortcomings in the original studies. In order to test these hypotheses, we searched patient records for occurrences of these variants at a premier genetic testing laboratory and reviewed the medical literature for initial ascertainment. We describe here a cautionary tale of broad relevance to genomic medicine.

METHODS

Study Populations

We used publicly-accessible sequence data from the NHBLI Exome Sequence Project (ESP),[11] 1000 Genomes Project (1000G),[14] and Human Genome Diversity Project (HGDP).[15] The NHBLI ESP has exome data from 4,300 European Americans and 2,203 African Americans; the 1000 Genomes Project Phase 1 has whole-genome data for 1,092 individuals from 14 worldwide populations; HGDP has whole-genome SNP data for 938 individuals from 51 worldwide populations. Clinical records for HCM patients were reviewed at the Laboratory for Molecular Medicine (LMM), Partners HealthCare, Boston, MA. All HCM patient reports with originally reported variant status "Pathogenic," "Presumed Pathogenic," "Unknown Significance," and "Pathogenicity Debated" were included (Table 2). The LMM patient population is a mixed population of 64% Caucasian and 8% black/African American individuals, with the remaining individuals of other or unspecified ancestry.[16]

Variant Ascertainment

A targeted search was performed for initial disease-variant associations for all HCM-associated high-frequency variants in the medical literature using PubMed. All

Human Genome Variation Society (HGVS) names for the variants (e.g., K247R and Lys247Arg) were used as well as all possible transcript variants obtained from NCBI dbSNP Build 140.[17] All original reports of disease-variant associations were in agreement with those listed in the Human Gene Mutation Database (HGMD) Version 2014.1.[13] "HCM-associated high-frequency variants" were defined as variants with minor allele frequency (MAF) greater than 1% in either NHLBI subpopulation.

Statistical and Bioinformatics Analyses

P values were computed using the chi-squared test. SNAP[18] was used to detect SNPs in linkage disequilibrium with high-frequency variants. The HGDP Selection Browser[19] was used to display allele frequencies in worldwide populations. The "penetrance" of a genetic variant is defined as the proportion of individuals with the variant who have HCM, expressed as the probability P(D|G) where D indicates the disease (HCM) and G indicates the variant:

$$\text{Penetrance} = P(D|G) = \frac{P(G|D)K}{P(G)}$$

The penetrance depends on the prevalence (K) as well as P(G|D), the proportion of HCM patients with the variant, and P(G), the overall frequency of the variant. Unless otherwise specified, all analyses were performed using the R statistical package.[20]

RESULTS

Only a few high-frequency variants account for the majority of HCM gene variation in the general population

The NHLBI Exome Sequence Project has previously been searched for any variant labeled a "Disease causing mutation" ("DM") for HCM in the Human Gene Mutation Database (HGMD Version 2012.2).[12] Although 94 distinct variants were discovered, we observed that relatively few variants account for the bulk of the genotype prevalence signal (Figure 1A). Five of the ninety-four HGMD HCM variants identified in the ESP data met our threshold to be "HCM-associated high-frequency variants" (MAF > 1% in either NHLBI subpopulation), and accounted for nearly 75% of the overall genotype prevalence signal.

HCM-associated high-frequency variants occur disproportionately in African Americans

All five HCM-associated high-frequency variants occurred at significantly greater frequencies in African Americans than in European Americans (Figure 1B, Chi-squared P < 0.001 for each comparison). The minor allele frequency for these five variants ranged from 1.5% to 14.9% in African Americans, 0.01% to 1.5% in European Americans, and 0.5% to 6.0% in the combined population. The genotype frequency, defined as (heterozygotes + homozygotes)/(total individuals), ranged from 2.9% to 27.1% in African Americans, 0.02% to 2.9% in European Americans, rand 1.0% to 11.1% in the combined population. The summed genotype frequency of the remaining 89 variants was not statistically different between African Americans and European Americans.

HCM-associated high-frequency variants have low penetrance in African Americans

We computed the penetrance for each variant across several clinical contexts (Supp. Figure 1). Because HCM occurs infrequently in the general population (Panel A) with a prevalence of K = 1:500,[2] even variants with minor allele frequency as small as 1% have a theoretical maximal penetrance of 0.2, but likely much lower, as the high allelic heterogeneity of HCM implies that $\mathrm{P}(G|D)$ is small for most variants, with a few notable exceptions.[21] Even large values of $\mathrm{P}(G|D)$ can have little influence on penetrance; for example, even if the *TNNT2* K247R variant were present in all African Americans with HCM, *K247R* would have a penetrance of less than 1%. Penetrance may take on rather different values in other clinical contexts (Panels B, C). Notably, for first-degree relatives, clinically insignificant high-frequency variants may have deceivingly large penetrance.

Clinically, all HCM-associated high-frequency variants are considered benign for all ethnicities

Using the clinical classification algorithm in use at the Laboratory for Molecular Medicine (Partners HealthCare Personalized Medicine),[7] we classified all high-frequency variants unambiguously as "benign," consistent with the LMM's current classification of these variants, given their elevated frequency in control populations as well as the mix of patient and functional data available for these variants. By contrast, in the HGMD database version 2014.1, four of the five variants remain classified in the most pathogenic category, "Disease causing mutation." Only one variant (*OBSCN* R4344Q) was downgraded from "disease-causing" to "disease-causing?" in September 2012.

HCM-associated high-frequency benign variants were classified as pathogenic in reports provided to African Americans

Seven patients, all of African or unspecified ancestry, received reports between 2005-2007 that one of the two benign variants *TNNI3* (P82S) or *MYBPC3* (G278E) was "Pathogenic" or "Presumed Pathogenic" (Table 2). In five of the seven reports, P82S or G278E was the most significant variant reported to the patient. Six additional inconclusive and positive cases reported later listed one of the two variants as "Unknown Significance" or "Pathogenicity Debated." Nine patients (of 13 total) had a clinical diagnosis of HCM, two had clinical features of HCM, and one had clinical symptoms of HCM. Five of 13 patients had a documented family history of HCM.

Small sample size and bias of original studies

All high-frequency variants were examined for their initial association in the medical literature (Table 1). For the two variants that affected patients, *TNNI3* P82S and *MYBPC3* G278E, control sample sizes were 85 and 100, which are below and equal to, respectively, the minimum currently accepted standards needed to corroborate pathogenicity.[4] Furthermore, none of the studies implicating these variants were undertaken in individuals of African ancestry explicitly; however, several studies might have sequenced or genotyped individuals of African ancestry during the discovery stage (Table 1). Generally, the original study that established the variant-HCM association consisted of three steps. First, HCM patients were sequenced at a handful of genes previously connected to the disease. Second, discovered variants were examined in ostensibly ethnically-matched unrelated controls and, where available, family members. Third, functional analyses were conducted in a subset of studies to assess causality of the variant.

African Americans have significantly more sequence variation than European Americans in both *MYBPC3* and *TNNI3*

We used the 1000G data to compare sequence diversity between African Americans and European Americans, using as proxies the populations ASW (Americans of African Ancestry in SW USA) and CEU (Utah Residents (CEPH) with Northern and Western European ancestry), respectively. As shown in Figures 2C and 2D, African Americans harbor significantly more segregating loci than European Americans in both genes. These "private sites," where MAF > 0% in one population but MAF = 0% in the other, are represented for ASW by the red points in Figures 2C and 2D. There are 66 (ASW) compared to 15 (CEU) private sites for *MYBPC3* and 45 (ASW) compared to 6 (CEU) private sites for *TNNI3*.

Diverse population sequence data reduce the risk of false positives

As shown in Figure 2B, even small studies of diverse populations are statistically well-powered to avoid misclassifying the five HCM-associated high-frequency variants. Conservatively, we used the lower frequency variant of the two that were misclassified in patients (*MYBPC3* G278E, MAF 0.0157 in African Americans, 0.000122 in European Americans). At these frequencies, even if African Americans constituted just 10% of the control cohort, we would have a 50% chance of correctly ruling out pathogenicity with a control cohort of only 200 individuals.

We documented how allele state and frequency for the HCM-associated high-frequency variants could be inferred both by neighboring variants (linkage disequilibrium, LD) and worldwide relationships (shared ancestry, admixture) (Figure 2a, Supp. Table 1). For example, the highest-frequency HCM variant (*TNNT2* K247R) was a locus in the Human Genome Diversity Project (HGDP)[15] (Figure 2A) and the HCM-associated high-frequency variant *TNNI3* P82S is in LD with the HGDP SNP rs7258659, which notably has non-zero allele frequencies in several African populations.

Paucity of available diverse control data may lead to the same errors in other populations

Table 3 shows the probability of ruling out pathogenicity for truly benign variants using existing sequencing resources. For example, using the 1000G population "Mexican Ancestry from Los Angeles" (MXL), which consists of 66 individuals, we have only a 1:2 chance of ruling out pathogenicity when the MAF is 0.5%. If MAF is 0.1%, such as for a rare variant discovered on high-coverage exome sequencing, the probability of ruling out pathogenicity is only 12% using the MXL population.

DISCUSSION

We hypothesized that high-frequency variants identified disproportionately in African Americans in the general population might have been previously misclassified in patients receiving genetic testing for hypertrophic cardiomyopathy. Upon reviewing patient records, we identified multiple individuals, all of African or unspecified ancestry, who had benign variants initially classified as pathogenic. Such misclassifications invalidate risk assessments undertaken in relatives, requiring a chain of amended reports and management plans, creating stress for patients and their families. Our findings suggest that false positive reports are an important and perhaps underappreciated component of the "genotype positive/phenotype negative" subset of tested individuals.[22]

To the best of our knowledge, this is the first illustration of how HCM variant reclassifications can disproportionately affect an underserved ethnic group. Consistent with previous work,[23] we observed significantly greater genetic diversity in African Americans in *MYBPC3* and *TNNI3*, the genes harboring G278E and P82S, respectively. When coupled with historically limited sequencing resources and bias in original studies, these findings suggest why African Americans might be disproportionately affected by variant reclassifications. Future work is needed to assess whether this pattern holds more broadly across other variants and types of misclassifications.

Minimizing misclassifications by sifting through genomic noise for causal variants is closely related to assessing penetrance, the proportion of individuals with the variant who express the disease. However, estimating penetrance is often difficult because it is sensitive to clinical context (Supp. Figure 1) and because many studies start with patients and ascertain variants as opposed to starting with the variant and prospectively evaluating patients and controls, a pattern not limited to HCM.[24] This approach is due, in part, to historically limited sequencing resources. Fortunately, recent large-scale sequencing efforts are mitigating this aspect of the variant annotation challenge,[11,25] while also introducing an unprecedented scale of novel variants and genes to consider.[6,26] While the NHLBI Exome Sequence Project is a powerful resource for African Americans and European Americans, a comparable resource for populations such as Native Americans and Asian Americans is urgently needed to prevent similar errors going forward (Table 3). Large-scale sequencing resources such as the NHLBI ESP are not only well-powered to "rule out" benign variants and reduce false positives (Figure 2B), but also allow pathogenicity to be corroborated for truly pathogenic variants (help "rule in" variants).

Large-scale sequencing data from the general population also enable systematic reassessments of prior disease-variant associations.[12,27,28] For such assessments in HCM, expert guidelines generally recommend using ethnically matched controls.[4] Doing so controls for false positives due to stratification provided the case and control ethnic mix is well matched. Ironically, insistence on using only ethnically matched controls may delay proper annotation if matching is imperfect. Consider *MYBPC3* G278E, an HCM-associated high-frequency variant that was discovered in a Parisian cohort[29] and misclassified in several African ancestry individuals (Table 2). Given the ethnic diversity of Paris and the fact that not all

HCM patients were of European origin in the original study (Table 1), it is conceivable that the discovery cohort included individuals of African ancestry. If only European-ancestry individuals were subsequently used as controls, then the study would have been underpowered to label the variant as non-pathogenic (Figure 2B). These findings suggest how current guidelines might be extended—variants from diverse ethnic groups may be used to rule out the pathogenicity of novel and known variants.[7,30] Such issues of population stratification are even subtler in admixed individuals, who have a patchwork of local ancestry[31] that defies the imperfect proxy of self-identified race.

Several steps are expected to improve care going forward. First, adopting a probabilistic framework alleviates much of the confusion in pathogenicity assessments because the quantities relevant to computing penetrance are incorporated explicitly and as continuous measures. Such a framework is required to achieve the infrastructural and statistical scaling challenges of "precision medicine."[32] Second, reevaluating the fragmented disease-variant literature depends on continued data-sharing and reporting standardization that are the aims of centralized databases like ClinVar.[33] Third, strengthening the relationship between the population genetics and medical genetics communities will lead to inventive safeguards against confounders like stratification. Lastly, as variant annotations are updated, an agile biomedical infrastructure would sense these changes and notify stakeholders expeditiously.[34] Indeed we expect that many "variants of uncertain significance"[6] will be recategorized in the near future as diverse control sequencing resources expand.

The cautionary tale we have described illustrates the complexities of variant classification. Far from being a clear binary decision, variant classification is an evolving art that will benefit most from a synergy of clinical, genetic, and statistical perspectives to prevent future misclassification errors and their adverse consequences.

REFERENCES

1. Maron, B.J. Hypertrophic Cardiomyopathy: A Systematic Review. *JAMA* **287**, 1308–1320 (2002).
2. Maron, B.J. et al. Prevalence of hypertrophic cardiomyopathy in a general population of young adults. Echocardiographic analysis of 4111 subjects in the CARDIA Study. Coronary Artery Risk Development in (Young) Adults. *Circulation* **92**, 785–789 (1995).
3. Maron, B.J. & Maron, M.S. Hypertrophic cardiomyopathy. *Lancet* **381**, 242–255 (2013).
4. Maron, B.J., Maron, M.S. & Semsarian, C. Genetics of Hypertrophic Cardiomyopathy After 20 Years: Clinical Perspectives. *J Am Coll Cardiol* **60**, 705–715 (2012).
5. Weidemann, F. et al. Long-term effects of enzyme replacement therapy on fabry cardiomyopathy: evidence for a better outcome with early treatment. *Circulation* **119**, 524–529 (2009).
6. Rehm, H.L. Disease-targeted sequencing: a cornerstone in the clinic. *Nat. Rev. Genet.* **14**, 295–300 (2013).
7. Duzkale, H. et al. A systematic approach to assessing the clinical significance of genetic variants. *Clin. Genet.* **84**, 453–463 (2013).
8. Richards, C.S. et al. ACMG recommendations for standards for interpretation and reporting of sequence variations: Revisions 2007. *Genet. Med.* **10**, 294–300 (2008).
9. Norton, N. et al. Evaluating pathogenicity of rare variants from dilated

cardiomyopathy in the exome era. *Circ Cardiovasc Genet* **5**, 167–174 (2012).

10. MacArthur, D.G. et al. A Systematic Survey of Loss-of-Function Variants in Human Protein-Coding Genes. *Science* **335**, 823–828 (2012).

11. *Exome Variant Server. Exome Variant Server* at <http://evs.gs.washington.edu/EVS/>

12. Andreasen, C. et al. New population-based exome data are questioning the pathogenicity of previously cardiomyopathy-associated genetic variants. *Eur J Hum Genet* **21**, 918–928 (2013).

13. Stenson, P.D. et al. Human Gene Mutation Database (HGMD): 2003 update. *Hum. Mutat.* **21**, 577–581 (2003).

14. McVean, G.A. et al. An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56–65 (2012).

15. Li, J.Z. et al. Worldwide human relationships inferred from genome-wide patterns of variation. *Science* **319**, 1100–1104 (2008).

16. Alfares, A. et al. Results of Clinical Genetic Testing of 2912 Probands with Hypertrophic Cardiomyopathy (submitted 2014).

17. Sherry, S.T. et al. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Research* **29**, 308–311 (2001).

18. Johnson, A.D. et al. SNAP: a web-based tool for identification and annotation of proxy SNPs using HapMap. *Bioinformatics* **24**, 2938–2939 (2008).

19. Pickrell, J.K. et al. Signals of recent positive selection in a worldwide sample of human populations. *Genome Res.* **19**, 826–837 (2009).

20. Ihaka, R. & Gentleman, R. R: A Language for Data Analysis and Graphics.

Journal of Computational and Graphical Statistics **5**, 299–314 (1996).

21. Dhandapany, P.S. et al. A common MYBPC3 (cardiac myosin binding protein C) variant associated with cardiomyopathies in South Asia. *Nat Genet* **41**, 187–191 (2009).

22. Maron, B.J., Yeates, L. & Semsarian, C. Clinical challenges of genotype positive (+)-phenotype negative (-) family members in hypertrophic cardiomyopathy. *Am. J. Cardiol.* **107**, 604–608 (2011).

23. Tishkoff, S.A. & Williams, S.M. Genetic analysis of African populations: human evolution and complex disease. *Nat. Rev. Genet.* **3**, 611–621 (2002).

24. Beutler, E., Felitti, V.J., Koziol, J.A., Ho, N.J. & Gelbart, T. Penetrance of 845G--> A (C282Y) HFE hereditary haemochromatosis mutation in the USA. *The Lancet* **359**, 211–218 (2002).

25. 1000 Genomes Project Consortium A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061–1073 (2010).

26. Keinan, A. & Clark, A.G. Recent explosive human population growth has resulted in an excess of rare genetic variants. *Science* **336**, 740–743 (2012).

27. Pugh, T.J. et al. The landscape of genetic variation in dilated cardiomyopathy as surveyed by clinical DNA sequencing. *Genet. Med.* (2014).doi:10.1038/gim.2013.204

28. Bick, A.G. et al. Burden of rare sarcomere gene variants in the Framingham and Jackson Heart Study cohorts. *Am. J. Hum. Genet.* **91**, 513–519 (2012).

29. Richard, P. et al. Hypertrophic cardiomyopathy: distribution of disease genes, spectrum of mutations, and implications for a molecular diagnosis strategy.

Circulation **107**, 2227–2232 (2003).

30. Ioannidis, J.P.A., Ntzani, E.E. & Trikalinos, T.A. "Racial" differences in genetic effects for complex diseases. *Nat Genet* **36**, 1312–1318 (2004).

31. Tang, H., Coram, M., Wang, P., Zhu, X. & Risch, N. Reconstructing genetic ancestry blocks in admixed individuals. *Am. J. Hum. Genet.* **79**, 1–12 (2006).

32. *Toward Precision Medicine: Building a Knowledge Network for Biomedical Research and a New Taxonomy of Disease.* 121 (National Academies Press: 2011).

33. Landrum, M.J. et al. ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Research* **42**, D980–5 (2014).

34. Wilcox, A.R. et al. A novel clinician interface to improve clinician access to up-to-date genetic results. *J Am Med Inform Assoc* **21**, e117–21 (2014).

35. García-Castro, M. et al. Hypertrophic cardiomyopathy: low frequency of mutations in the beta-myosin heavy chain (MYH7) and cardiac troponin T (TNNT2) genes among Spanish patients. *Clin. Chem.* **49**, 1279–1285 (2003).

36. Arimura, T. et al. Structural analysis of obscurin gene in hypertrophic cardiomyopathy. *Biochem. Biophys. Res. Commun.* **362**, 281–287 (2007).

37. Niimura, H. et al. Sarcomere protein gene mutations in hypertrophic cardiomyopathy of the elderly. *Circulation* **105**, 446–451 (2002).

38. Matsushita, Y. et al. Mutation of junctophilin type 2 associated with hypertrophic cardiomyopathy. *J. Hum. Genet.* **52**, 543–548 (2007).

39. Gravel, S. et al. Reconstructing Native American migrations from whole-genome and whole-exome data. *PLoS Genet.* **9**, e1004023 (2013).

40. Friedlaender, J.S. et al. The genetic structure of Pacific Islanders. *PLoS Genet.* **4**, e19 (2008).

Table 1: Studies that initially implicated HCM-associated high-frequency variants

Gene (Variant)	*Reference*	Discovery	Controls	*In vitro*	*In vivo*	Country	LMM Clinical Panel
TNNT2 (K247R)	Garcia-Castro (2003)[35] *Clin Chem* 49, 1279	Targeted gene sequencing of unrelated cases (Asturias)	200 (Asturias)	-	-	Spain	+
*OBSCN** *(R4344Q)*	Arimura (2007)[36] *Biochem Biophys Res Commun* 362,281	Targeted gene sequencing of unrelated cases (Japanese)	288 (Japanese)	+	-	Japan	-
TNNI3 (P82S)	Niimura (2002)[37] *Circulation* 105, 446	Targeted gene sequencing of unrelated cases**	85**	-	-	USA	+
MYBPC3 (G278E)	Richard (2003)[29] *Circulation* 107, 2227	Targeted gene sequencing of unrelated cases***	100***	-	-	France	+
JPH2 (G505S)*	Matsushita (2007)[38] *J Hum Genet* 52, 543	Targeted gene sequencing of cases (Japanese)	236 (Japanese)	+	-	Japan	-

* *OBSCN* and *JPH2* have never been included in cardiomyopathy testing at the LMM.
** No specific ethnicity provided, but "informed consent was obtained in accordance with human subject committee guidelines at Brigham and Women's Hospital, St. George's Hospital Medical School [U.K.], and Minneapolis Heart Institute Foundation."
*** "Patients were recruited in France, and most of them were of European origin."

Table 2: Clinical findings for HCM-associated high-frequency variants

Age	Ethnicity	Report Year	Report	Variant	Originally Reported Status	Current Status	Most Significant?	Indication for Test
46	Unavailable	2005	Positive	Pro82Ser	Pathogenic	Benign	Y	Clinical Diagnosis of HCM
75	Unavailable	2005	Positive	Pro82Ser	Pathogenic	Benign	Y	Family History and Clinical Symptoms of HCM
32	Black or African American	2005	Positive	Pro82Ser	Presumed Pathogenic	Benign	N	Clinical Diagnosis of HCM
34	Black or African American	2005	Positive	Pro82Ser	Pathogenicity Debated	Benign	N	Clinical Diagnosis and Family History of HCM
12	Black or African American	2006	Inconclusive	Pro82Ser	Unknown Significance	Benign	Y	Family History of HCM
40	Black or African American	2007	Inconclusive	Pro82Ser	Unknown Significance	Benign	Y	Clinical Diagnosis of HCM
45	Black or African American	2007	Inconclusive	Pro82Ser	Unknown Significance	Benign	Y	Clinical Features of HCM
16	Asian	2008	Positive	Pro82Ser	Unknown Significance	Benign	N	Clinical Diagnosis and Family History of HCM
59	Black or African American	2006	Positive	Gly278Glu	Presumed Pathogenic	Benign	Y	Clinical Features of HCM
15	Black or African American	2007	Positive	Gly278Glu	Presumed Pathogenic	Benign	Y	Clinical Diagnosis of HCM
16	Black or African American	2007	Positive	Gly278Glu	Presumed Pathogenic	Benign	Y	Clinical Diagnosis of HCM
22	Black or African American	2007	Positive	Gly278Glu	Presumed Pathogenic	Benign	N	Clinical Diagnosis and Family History of HCM
48	Black or African American	2008	Positive	Gly278Glu	Unknown Significance	Benign	N	Clinical Diagnosis of HCM

*The "Most Significant?" column indicates whether the variant was unequivocally the most pathogenic variant on the original report provided to the patient.

Table 3: Limited control sequencing resources for ruling out pathogenicity in several US populations

U.S. Census Category	Cohort	Proxy Population	N	MAF = 0.5%	MAF = 0.1%	MAF = 0.01%
White	NHLBI ESP	European Americans	4400	100%	100%	22%
Black or African American	NHLBI ESP	African Americans	2203	100%	99%	36%
American Indian and Alaska Native	1000 Genomes	Mexican Ancestry from Los Angeles, USA[39]	66	48%	12%	1%
Asian	1000 Genomes	Han Chinese in Beijing, China	97	62%	18%	2%
Native Hawaiian and Other Pacific Islander	1000 Genomes	Southern Han Chinese[40]	100	63%	18%	2%

The probability of ruling out pathogenicity is shown for different self-identified ethnicities using proxy populations.

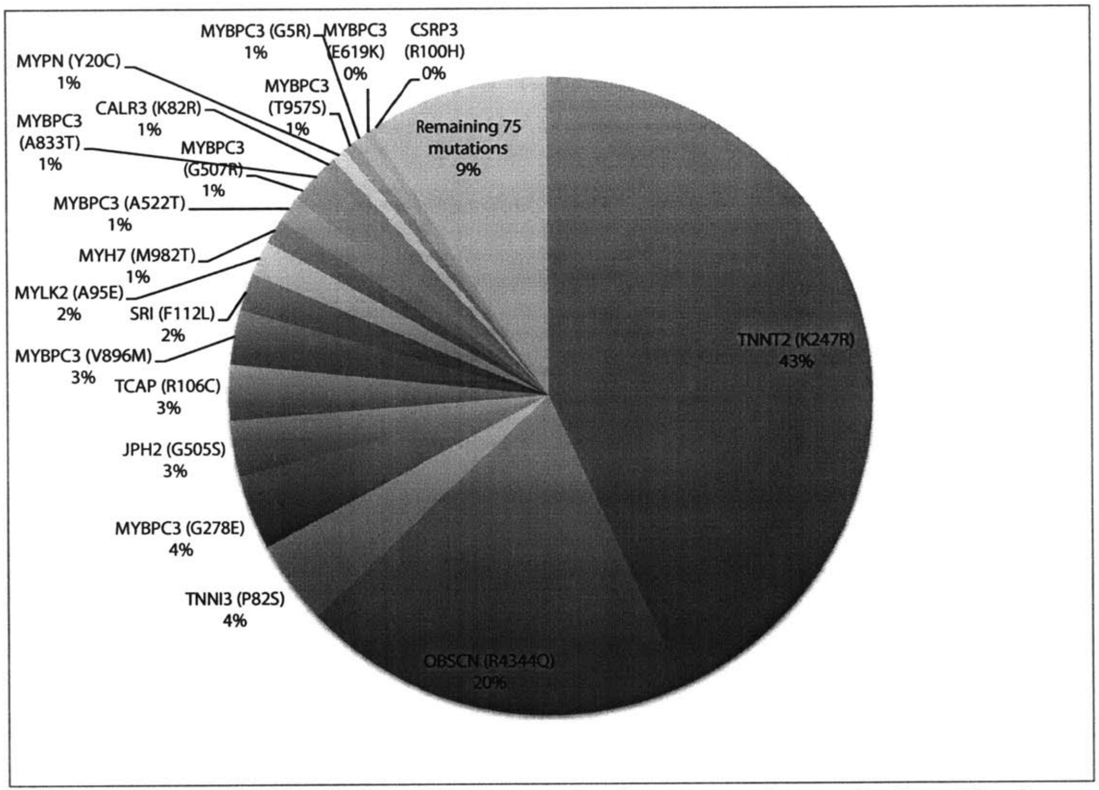
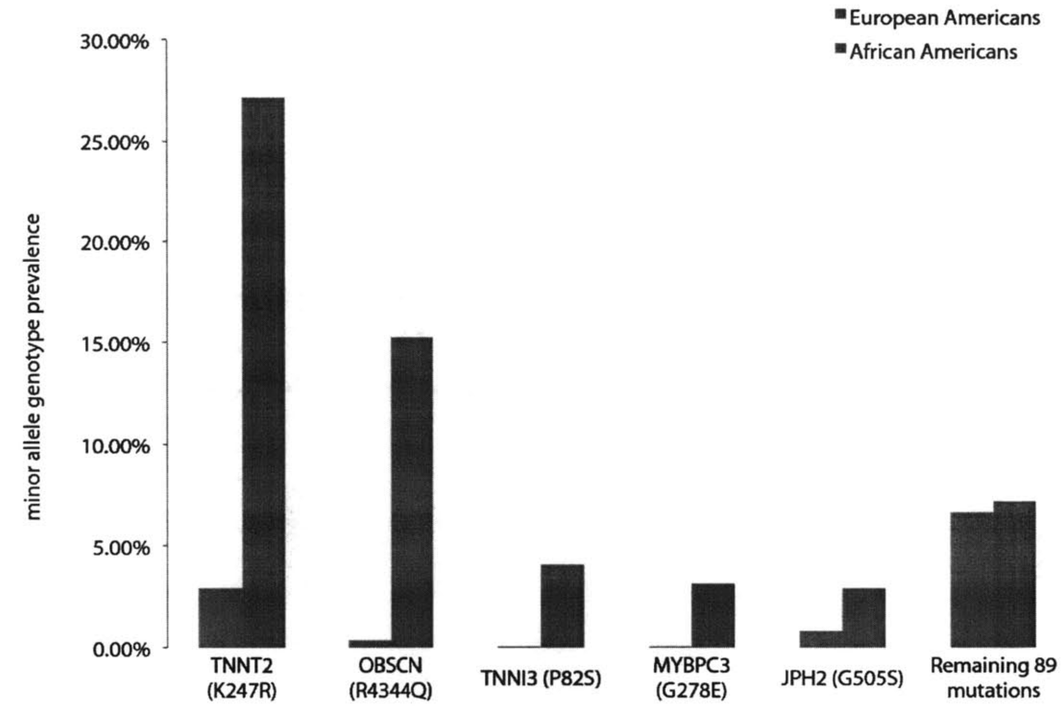


Figure 1A: Overrepresented HCM variants in the general population. The five highest-frequency variants account for 74% of the genotype frequency signal for HCM in the general population.



	TNNT2 (K247R)	OBSCN (R4344Q)	TNNI3 (P82S)	MYBPC3 (G278E)	JPH2 (G505S)	Remaining 89 mutations
European Americans	2.88%	0.33%	0.03%	0.02%	0.80%	6.67%
African Americans	27.14%	15.27%	4.07%	3.15%	2.92%	7.18%

Figure 1B: All HCM-associated high-frequency variants are significantly more common in African Americans than European Americans. Chi-squared P < 0.001 for each comparison.

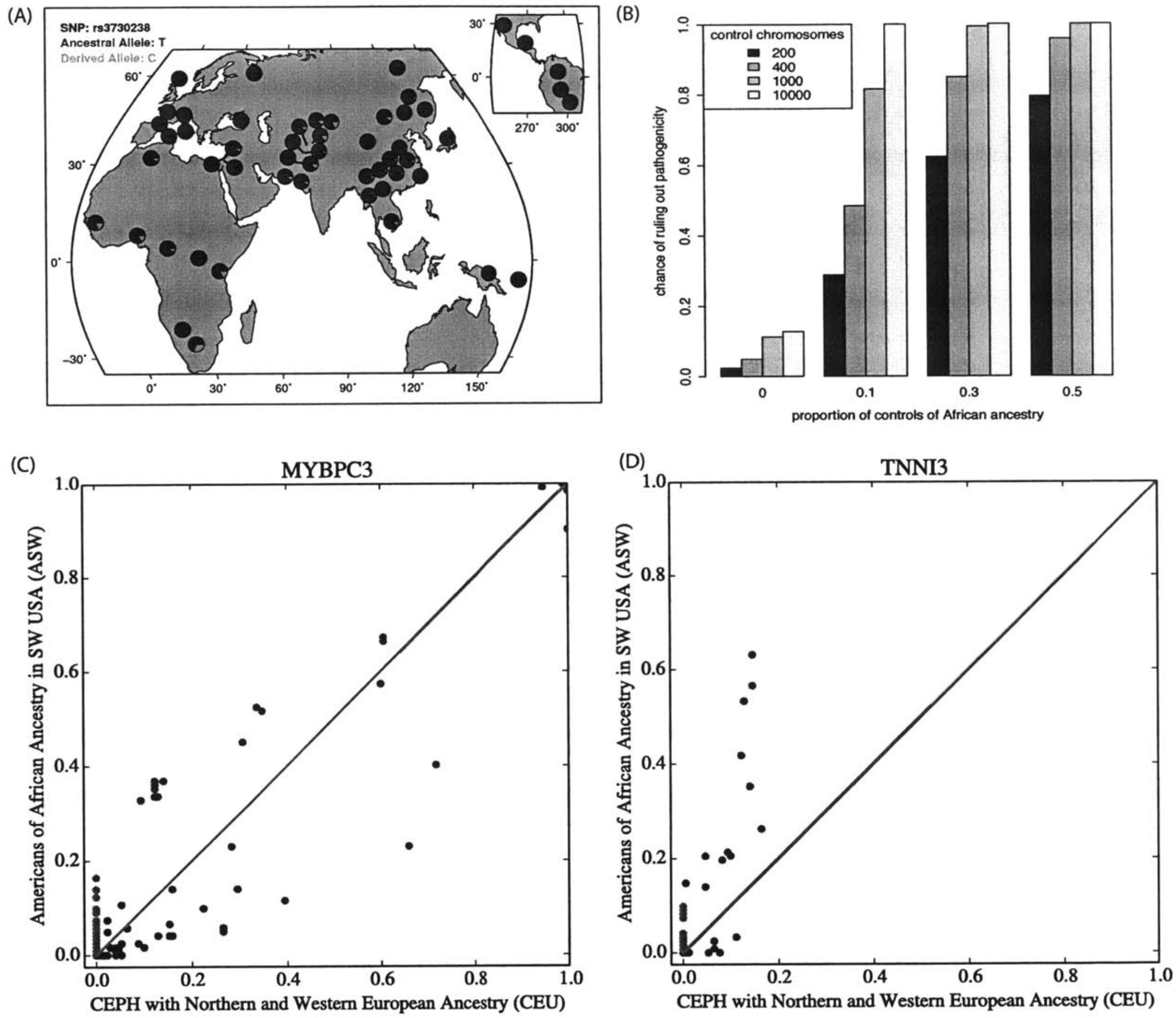
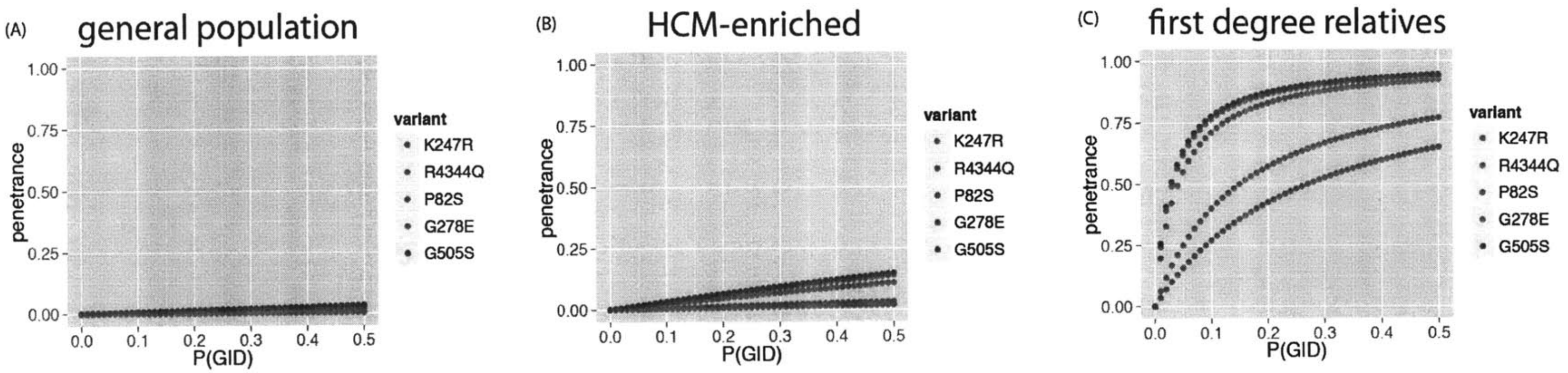


Figure 2: Diverse sequencing data help prevent variant misclassifications by illuminating hidden sources of bias as well as useful correlations to infer allele frequency (A) *TNNT2* (K247R) was a variant genotyped in the HGDP. Most populations around the world have non-zero minor allele frequency. **(B)** For a variant predominantly found in one ethnic group, the chance of correctly ruling out pathogenicity for a truly benign variant generally increases with the diversity of the control cohort and the number of controls (control chromosomes shown in legend). These simulations use the allele frequencies of the *MYBPC3* G278E variant, which has an African American minor allele frequency (MAF) of 0.0157 and a European American MAF of 0.000122. **(C/D)** MAFs for 1000G populations ASW (y-axis, 61 individuals) and CEU (x-axis, 85 individuals) for the HCM genes *MYBPC3* and *TNNI3*. Each point represents a distinct variant (SNP/indel). African Americans have significantly more private variants (CEU MAF = 0% and ASW MAF > 0%, colored in red), than European Americans (ASW MAF = 0% and CEU MAF > 0%).

Supplementary Table 1: Allele frequencies in global populations

			African			European						
Gene (Variant)	**AA**	**EA**	**ASW**	**YRI**	**LWK**	**CEU**	**IBS**	**GBR**	**TSI**	**FIN**	**JPT**	**dbSNP**
TNNT2 (K247R)	14.89%	1.48%	11.5%	17.6%	11.9%	0.6%	3.6%	1.7%	2.6%	4.3%	7.3%	rs3730238
OBSCN (R4344Q)	7.97%	0.17%	10.7%	15.9%	17.0%	0%	0%	1.1%	0%	0%	0%	rs79023478
TNNI3 (P82S)	2.03%	0.01%	3.3%	1.7%	0%	0%	0%	0%	0%	0%	0%	rs77615401
MYBPC3 (G278E)	1.57%	0.01%	0.8%	1.1%	2.1%	0%	0%	0%	0%	0%	0%	rs147315081
JPH2 (G505S)	1.46%	0.40%	3.3%	3.4%	3.6%	0%	0%	0.6%	3.1%	2.2%	0%	rs140740776
	NHLBI Exome Sequence Project		1000 Genomes Project									

Minor allele frequencies in populations around the world for the five HCM-associated high-frequency variants. African populations include ASW (Americans of African Ancestry in SW USA), YRI (Yoruba in Ibadan, Nigeria), and LWK (Luhya in Webuye, Kenya). European populations include CEU (Utah Residents (CEPH) with Northern and Western European ancestry), IBS (Iberian population in Spain), GBR (British in England and Scotland), TSI (Toscani in Italia), and FIN (Finnish in Finland). Also shown are minor allele frequencies for JPT (Japanese in Tokyo, Japan).



Supplementary Figure 1: Computed penetrance of HCM-associated high-frequency variants in three different clinical contexts. (A) General population (prevalence K = 1:500), **(B)** population enriched for HCM patients, e.g., mixed population of HCM patients and general population (K = 1:100), **(C)** first-degree relatives (K = 1:2). P(G|D) is the proportion of HCM patients with the variant.

This page is intentionally left blank.

Chapter 3

The precarious wisdom of communal science

Overview

Shared datasets receive little attention with regards to reproducibility. Broader reproducibility initiatives[1,2] focus on investigator responsibility, but structural aspects of the scientific enterprise including a shared dataset's access policy and the distribution of studies across relationships are equally influential, yet are neither controlled by nor visible to individual researchers. Here we introduce a new quantity called the "dataset positive predictive value" (*dPPV*) to quantify the proportion of true claims amongst all claims made during multiple uses of a shared dataset by different investigators. We show that, in the presence of moderate bias, using nominal statistical significance levels to make claims leads to publications that are mostly false. The distribution of studies performed across relationships affects both the number of claimed relationships and the proportion that is true, even when teams use identical data and inference procedures. We derive scaling rules that hold generally and demonstrate several surprising facts about the reproducibility of communal science. For example, reproducibility often declines as more teams study the same topic and small pilot studies may produce more confusion than guidance. Finally, we discuss data access policies. We find that restrictive data access policies may blur evaluations of reproducibility more than open policies. We discuss possible solutions to prevent confusion and maintain reproducibility amid the competing interests of communal science.

INTRODUCTION

Shared data resources permit a multiplicity of uncoordinated investigations to be performed. This multiplicity rivals high-throughput omics multiple hypothesis-testing in scale but not synchrony: whereas high-throughput omics analyses often use a single test platform where all measurements are taken simultaneously, analyses using shared data may happen piece-meal by investigators over decades.[3] This pattern of use is likely to become more common as shared datasets and user-centered data browsers[4] flourish. Failure to account for these multiple tests, especially in the context of biases like selective reporting, can lead to false[5] or inflated[6] claims. In order to enhance the reproducibility of communal science, we develop an analytical framework for a new quantity called the "dataset positive predictive value" (*dPPV*) to measure the proportion of true claims amongst all claims made during multiple uses of a shared dataset. We use this framework to understand the reliability of claims made during multiple uses of a shared dataset by using simulations that vary which relationships are tested and policies that govern database access and data sharing.

METHODS

dPPV quantifies reproducibility in the context of a shared dataset's access policy and features of the studies that are performed (Box, Figure 1). The actors in this framework are research teams, the relationships they choose to study, and the policies governing data usage. Capturing the full range of shared dataset use requires extending previous models for scientific reproducibility[5] in two key ways.

First, we allow any team to study any relationship in the data over time, as opposed to focusing on the more limited scenario where all teams are testing the same relationship[5]. Second, we formally model the variance in realized values of *dPPV*, which is due to a mixture of luck, study features, and the truth of relationships tested.

Consider *c* relationships of which c_T are "non-null" and c_F are "null" relationships, under study by *n* teams at time *t*. Teams may be studying one or more relationships, with the full set of ongoing studies specified by the dynamic *n* x *c* binary matrix *N* (Box). Each study is performed at significance level α, and team *i* is studying relationship *j* with Type II Error β_{ij}. The ratio of "non-null" to "null" relationships is given by *R*. The bias term, *u*, is the proportion of research findings that would not have been claimed under ideal study design, analysis and reporting procedures[5]. Tying together these terms, *dPPV*(*t*) is the proportion of claimed relationships that is true at time *t*, and can be written as the ratio of two Poisson binomial random variables:

$$dPPV(t) = \frac{\mathrm{PB}(c_T(t), p_j(t))}{\mathrm{PB}(c_T(t), p_j(t)) + \mathrm{PB}(c_F(t), q_j(t))} \qquad (1)$$

where $p_j(t)$ and $q_j(t)$ are probability vectors that describe the features of studies underway at time *t* and are based on β_{ij} and α, respectively, as well as the extent of bias *u* (Box).

We developed a simulation framework to study *dPPV* (Figure 1). We obtained laboratory data (e.g. serum creatinine and glucose) of participants in the National Health and Nutrition Examination Survey (NHANES),[7] a cross-sectional epidemiological study. Next, we created a large set of uncorrelated "synthetic variables" by randomly sampling from the empirical distributions of the NHANES analytes. Synthetic variables were combined in varying proportions with the NHANES laboratory data for each patient, allowing control over the ratio of "non-null" to "null" relationships (R) as detailed below. Pairwise relationships between variables in the combined data were measured with the Pearson product-moment correlation coefficient, computed after adjusting appropriately for the NHANES sampling design[8]. In our simulation, a "non-null" relationship is defined as any pairwise correlation between NHANES analytes that met a Bonferroni-adjusted significance threshold ($p < 1.8 \times 10^{-4}$). All other pairwise correlations, whether involving only NHANES analytes or synthetic variables, are "null". During each time step of the simulation, several studies are approved, performed, selected and research findings are claimed if they meet a threshold of statistical significance (a model parameter). We track the "true" proportion of claimed relationships over time in a dataset in which a proportion of the potentially testable correlations are simulated as being "null" (Figure 1).

RESULTS

Statistically significant correlations that represent false-positives are readily observed between artificial variables and the NHANES clinical variables. If research teams conduct uncoordinated investigations of separate relationships and each team uses a nominal statistical significance threshold ($p < 0.05$), in expectation there will be 0.05 x ($c/(R+1)$) false claimed relationships. For example, consider a shared dataset with relatively few true relationships (R = 0.001) but many ongoing studies (c = 1000). If the ongoing studies are not somehow enriched for targeting a higher proportion of true relationships and always have R=0.001, we expect 50 statistically significant false positive discoveries and just 1 true positive discovery assuming reasonably good power ($1 - \beta = 0.8$), yielding an expected *dPPV* of less than 2%. While the nominally acceptable power exceeds the false positive rate by a factor of 16, the null relationships dwarf the non-null relationships by a factor of 1000, dooming *dPPV*.

The expected value of *dPPV* can be computed more generally:

$$\mathrm{E}(dPPV) \approx \frac{\mu_{\mathrm{TP}}}{\mu_{\mathrm{TP}} + \mu_{\mathrm{FP}}} \qquad (2)$$

where μ_{TP} and μ_{FP} are the expected number of true positives and false positives, respectively. Intuitively, the expected proportion of true relationships claimed from communal use of a shared dataset is the ratio of the expected number of true positives to the expected number of total claimed relationships. While an analogous formula holds for the model in which research teams study separate relationships with identical power and statistical significance thresholds,[5] it is noteworthy that a similar relationship holds in the more complicated scenario of many teams separately mining a shared dataset (Chapter 4).

Equation 2 describes how reproducibility declines during communal science scenarios that are likely to happen in real life (Table 1). Consider the situation in which teams arriving late to a shared dataset choose to study relationships with reduced pre-study odds (i.e. addressing more far-fetched relationships) or with reduced power (e.g. addressing more rare phenomena). For example, most relationships with high *a priori* probability (low-hanging fruit) may have already been claimed and to study hypotheses without precedent, late-arriving researchers may study *a priori* less likely hypotheses. Alternatively, late-arriving teams may choose to investigate potential subgroup effects across different strata or relationships that have smaller numbers of observations. It is well documented in the clinical trials literature that such post-hoc analyses ought to be treated with caution.[9,10] An analogous result holds for communal use of a shared dataset (Figure 3a). Some additional common scenarios are listed in Table 1.

Another critical parameter is bias (u), which may arise for a variety of reasons including, but not limited to, conflicts of interest,[11] heterogeneity,[12] confounding,[13] questionable research practices,[14] or inappropriate analysis procedures.[6] Analysis challenges may spread as big data is commoditized and may worsen as data creators and data analysts become increasingly disconnected. For example, researchers analyzing administrative claims[15] or electronic health record (EHR) data (e.g. comparative effectiveness), or researchers monitoring treated patients for adverse events (e.g. pharmacovigilance), are often disconnected from the individuals who organize, code, and store these data. Moreover, many of these datasets were never created for research purposes and there may be a poor understanding by their users of the analytical caveats resulting from poor quality and deficiencies in the data and measurements available. Unless proper analysis guidelines are maintained or techniques such as pre-specified falsification end points are used,[16] reproducibility is liable to suffer when bias (u) increases in these settings (Figure 3c).

More optimistically, Equation 2 reveals how to improve reproducibility in communal science. For example, ensuring adequate power lessens the false discovery risks from communal pursuit of less likely hypotheses (Figure 3a). Manipulating *dPPV* shows that the pre-study odds and dataset-averaged power are closely related to one another. If the equation

$$(R_{\text{eff}}) \times (1/c_T \sum_{j=1}^{c_T} 1 - [1 - u]\Pi_{i=1}^{n} \beta_{i,j}^{n_{i,j}}) = \text{constant} \tag{3}$$

holds, the expected value of *dPPV* is unchanging, assuming the false positive rate remains constant. In other words, doubling the relationship-averaged power compensates for studying hypotheses half as likely. For example, when using EHR data to study rare diseases, combining patient cohorts across hospitals[17] may enable the community to preserve reproducibility, partially offsetting a “file drawer” problem[18] that may worsen as increased sharing of large data sets is promoted. However, the same resource may be used to justify almost any statistically significant yet clinically insignificant effect with an appropriately large sample size (e.g. even if the true relative risk is 1.001, a sufficiently large population will achieve statistical significance).

Reproducibility also improves if the community is able to limit certain forms of "herding."[19] Herding occur when an initial discovery is followed by a wave of similar studies, either by the same team or others. In the context of a shared dataset, this behavior may take the form of different groups analyzing the same topic using slices of data from the larger shared dataset. Empirical evidence also suggests that when the same data are re-analyzed, even with participation of the original authors, different conclusions are often reached, even for non-exploratory research such as randomized trials.[20] Independent studies of the same relationship cause the probability of a chance false discovery to rise rapidly, with biases like selective reporting and nonstandard analysis procedures worsening the situation (Figure 3c). This pattern may seem counterintuitive, as corroboration by independent teams is considered to be the hallmark of scientific progress.

Independent corroboration is necessary to verify or refute claimed relationships during the *replication* stage. But during the *discovery* stage, when novel relationships are being claimed and research teams can select which findings to publish, the presence of many teams studying the same relationship threatens reproducibility because the chance of a false discovery rises steeply with the number of teams. Using *dPPV*, the expected rate of false claimed relationships is given by:

$$1/c_F \sum_{j=1}^{c_F} 1 - \Pi_{i=1}^{n}(1-\alpha)^{n_{i,j}} \tag{4}$$

where the sum spans the false relationships under study. Given the same total number of studies, reducing the number of studies per relationship tends to reduce the expected rate of false claims (Figure 3d, Chapter 4). This means that for discovery research, we should think of strategies that reduce the chances that different teams would be pursuing the same or overlapping research questions. Intellectual crowding is bad for discovery.

One discussed strategy to narrow the scope of hypotheses considered is to conduct small pilot studies and follow up only promising findings. Many funding agencies have incentives for small pilot studies and grant applications routinely request pilot data, hoping to avoid committing large funds and resources in areas where there is yet no evidence that they would be fruitful. However, it is not well known how this strategy influences reproducibility in the setting of shared datasets. The matrix β(t) in *dPPV* stores the Type II Error of all studies, so constrained time evolution of *dPPV* simulates pilot studies that tend to have smaller sample sizes and reduced power. In general, simulations show that small pilot studies tend to increase the expected value of *dPPV* (Figure 3b). However, here the expected value of *dPPV* may not suffice to convey the entire picture. Very small pilot studies are especially concerning because there is larger unpredictability in what they will find: then they may lead to overreaction, either by prematurely terminating a research program or justifying undue investment.[21] To capture the extent of this potential problem, we need to assess the variance of the *dPPV*, which is higher when smaller studies are performed.

We have focused primarily on the *expected value* of *dPPV*, but the proportion of true claimed relationships takes on different values even with identical reproducibility parameters (Figure 4). This *variance* is relevant in practice because it can lead to confusion and divert attention to factors unrelated to reproducibility. Variance depends on several factors that are not controlled by individual teams: these include the data governance policy and the number of teams simultaneously mining the dataset. These dependencies are captured in part by the means μ_{TP} and μ_{FP} in the variance equation for *dPPV*:

$$\mathrm{Var}[dPPV(t)] \approx \frac{\mu^2_{\mathrm{TP}}}{\mu^2_{\mathrm{TP+FP}}}\left(\frac{\mathrm{Var}(TP)}{\mu^2_{\mathrm{TP}}} + \frac{\mathrm{Var}(TP+FP)}{\mu^2_{\mathrm{TP+FP}}} - 2\frac{\mathrm{Cov}(TP, TP+FP)}{\mu_{\mathrm{TP}}\mu_{\mathrm{TP+FP}}}\right) \tag{5}$$

In general, more liberal data governance policies reduce variance (Figure 4a). If a large number of relationships are studied, estimation of *dPPV* improves due to the law of large numbers and as can be seen in Equation 5 (Chapter 4). By extension, reproducibility tracks with the ebbs and flows of scientific interest or funding in a particular field (Figure 4b). Chance is at work in both of these settings—understanding the degree of its influence will temper expectations and avoid misattribution of both blame and credit for efforts designed to enhance reproducibility.

DISCUSSION

Most recent efforts to enhance reproducibility emphasize the responsibility of individual researchers.[1,2] However, *dPPV* shows that there are structural factors such as the data governance policy that can be equally influential. These factors affect both the *expectation* and the *variance* of *dPPV*. Understanding the spectrum of possible values of *dPPV* and similar metrics is a prerequisite for systematically improving reproducibility. Estimation of *dPPV* demonstrates that such assessments are laden with predictable amounts of variability that worsen as evaluations become narrower—the global evidence is more stable than the evidence procured by single institutions or departments, which is more stable than the evidence procured by individual investigators. An institution or data governing body might initially adopt small reproducibility programs to audit a handful of studies in order to closely monitor reproducibility. Ironically, such initiatives may be the least reliable. In the current globalized environment for research, single academic or research institutions may have access to very thin slices of the total evidence, and pilot programs within these thin slices may be particularly unreliable and have very high variance.

dPPV specifies how to assess and improve reproducibility in communal science. The principal challenge in estimating and applying *dPPV* to existing shared datasets is that model parameters such as u, number of hypotheses tested, and R may be unavailable or unmeasured. The relationships we have described hold generally, though their magnitude will be sensitive to specific parameter values. However, very often we may have empirical evidence on the potential range of bias and pre-study odds in different fields. Moreover, even subjective parameter assessments are likely to improve the scientific process and help contextualize novel claimed relationships.[22] But most promising would be to use the wealth of rich shared data to estimate empirically the reproducibility parameters across a wide range of communal science pursuits (Table 1). This is likely to be a trial-and-error, iterative process.

Estimates of *dPPV* model parameters can be obtained in both agnostic[23-25] and hypothesis-driven approaches to discovery. For example, genome-wide association studies (GWAS)[26,27] offer a paradigm where reproducibility has been high.[28] It is now standard in GWAS for researchers to correct for multiple hypotheses, minimize bias from ancestry-based confounding,[29] and even estimate the pre-study odds using concepts such as heritability.[30] If separate researchers tested individual SNPs or genes instead of utilizing GWAS methodology, irreproducibility and wasted investment would be catastrophically high, as well documented for the thousands of claims made by single investigators in the candidate gene era.[31] Systematic methods of agnostic inquiry have been applied to other fields[24] but remain a challenge for the majority of observational clinical data, reflecting in part the siloed efforts that characterize much of biomedical research over the past several decades. The successful scientist is still considered to be the individualistic principal investigator who comes up with extravagant discoveries against the prior odds. Paradoxically, this scientist model is linked to very low reproducibility. Conversely, not a single Nobel Prize has been given to communal, team science to-date, despite its tremendous—and growing—importance in biomedicine and despite the fact that it is a model that is likely to be associated with high reproducibility.

Without appreciating the influence of structural factors including a shared dataset's access policy and the distribution of studies across relationships, we have only an incomplete understanding of the risks of irreproducibility and the levers with which we can increase the reproducibility of communal science. As the scientific community strives to enhance reproducibility, successful efforts will require methods to overcome the fickle nature of scientific reproducibility in the context of today's communal science.

REFERENCES

1. Collins, F.S. & Tabak, L.A. Policy: NIH plans to enhance reproducibility. *Nature* **505**, 612–613 (2014).

2. McNutt, M. Reproducibility. *Science* **343**, 229–229 (2014).

3. Colditz, G.A., Manson, J.E. & Hankinson, S.E. The Nurses' Health Study: 20-year contribution to the understanding of health among women. *J Womens Health* **6**, 49–62 (1997).

4. Athey, B.D., Braxenthaler, M., Haas, M. & Guo, Y. tranSMART: An Open Source and Community-Driven Informatics and Data Sharing Platform for Clinical and Translational Research. *AMIA Jt Summits Transl Sci Proc* **2013**, 6–8 (2013).

5. Ioannidis, J.P.A. Why most published research findings are false. *PLoS Med.* **2**, e124 (2005).

6. Ioannidis, J.P.A. Why most discovered true associations are inflated. *Epidemiology* **19**, 640–648 (2008).

7. *National Center for Health Statistics (NCHS). National Health and Nutrition Examination Survey Questionnaire (or Examination Protocol, or Laboratory Protocol). National Center for Health Statistics (NCHS). National Health and Nutrition Examination Survey Questionnaire (or Examination Protocol, or Laboratory Protocol)* (Centers for Disease Control and Prevention (CDC).:).at <http://www.cdc.gov/nchs/nhanes/>

8. Lumley, T. Analysis of complex survey samples. *Journal of Statistical Software* **9**, 1–19 (2004).

9. Pocock, S.J., Assmann, S.E., Enos, L.E. & Kasten, L.E. Subgroup analysis, covariate adjustment and baseline comparisons in clinical trial reporting: current practice and problems. *Stat Med* **21**, 2917–2930 (2002).

10. Thompson, S.G. & Higgins, J.P.T. How should meta-regression analyses be undertaken and interpreted? *Stat Med* **21**, 1559–1573 (2002).

11. Bekelman, J.E., Li, Y. & Gross, C.P. Scope and impact of financial conflicts of interest in biomedical research: a systematic review. *JAMA* **289**, 454–465 (2003).

12. Madigan, D. et al. Evaluating the impact of database heterogeneity on observational study results. *Am. J. Epidemiol.* **178**, 645–651 (2013).

13. Smith, G.D. et al. Clustered environments and randomized genes: a fundamental distinction between conventional and genetic epidemiology. *PLoS Med.* **4**, e352 (2007).

14. Fanelli, D. How many scientists fabricate and falsify research? A systematic review and meta-analysis of survey data. *PLoS ONE* **4**, e5738 (2009).

15. Ioannidis, J.P.A. Are Mortality Differences Detected by Administrative Data Reliable and Actionable? *JAMA* **309**, 1410–1411 (2013).

16. Prasad, V. & Jena, A.B. Prespecified falsification end points: can they validate true observational associations? *JAMA* **309**, 241–242 (2013).

17. Murphy, S. et al. Instrumenting the health care enterprise for discovery research in the genomic era. *Genome Res.* **19**, 1675–1681 (2009).

18. Rosenthal, R. The file drawer problem and tolerance for null results. *Psychological Bulletin* **86**, 638–641 (1979).

19. Young, N.S., Ioannidis, J.P.A. & Al-Ubaydli, O. Why current publication practices may distort science. *PLoS Med.* **5**, e201 (2008).

20. Ebrahim, S. et al. Reanalyses of Randomized Clinical Trial Data. *JAMA* **312**, 1024–1032 (2014).

21. Loscalzo, J. Pilot Trials in Clinical Research: Of What Value Are They? *Circulation* **119**, 1694–1696 (2009).

22. Patel, C.J. & Ioannidis, J.P.A. Placing epidemiological results in the context of multiplicity and typical correlations of exposures. *J Epidemiol Community Health* (2014).doi:10.1136/jech-2014-204195

23. Visscher, P.M., Brown, M.A., McCarthy, M.I. & Yang, J. Five years of GWAS discovery. *Am. J. Hum. Genet.* **90**, 7–24 (2012).

24. Patel, C.J., Bhattacharya, J. & Butte, A.J. An Environment-Wide Association Study (EWAS) on type 2 diabetes mellitus. *PLoS ONE* **5**, e10746 (2010).

25. Patel, C.J. & Ioannidis, J.P.A. Studying the elusive environment in large scale. *JAMA* **311**, 2173–2174 (2014).

26. Gibbs, R.A. et al. The International HapMap Project. *Nature* **426**, 789–796 (2003).

27. 1000 Genomes Project Consortium A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061–1073 (2010).

28. Visscher, P.M., Brown, M.A., McCarthy, M.I. & Yang, J. Five Years of GWAS Discovery. *The American Journal of Human Genetics* **90**, 7–24 (2012).

29. Price, A.L. et al. Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* **38**, 904–909 (2006).

30. Visscher, P.M., Hill, W.G. & Wray, N.R. Heritability in the genomics era — concepts and misconceptions. *Nat. Rev. Genet.* **9**, 255–266 (2008).

31. Ioannidis, J.P.A., Tarone, R. & McLaughlin, J.K. The False-positive to False-negative Ratio in Epidemiologic Studies. *Epidemiology* **22**, 450–456 (2011).

32. Takahashi, K. & Yamanaka, S. Induction of Pluripotent Stem Cells from Mouse Embryonic and Adult Fibroblast Cultures by Defined Factors. *Cell* **126**, 663–676 (2006).

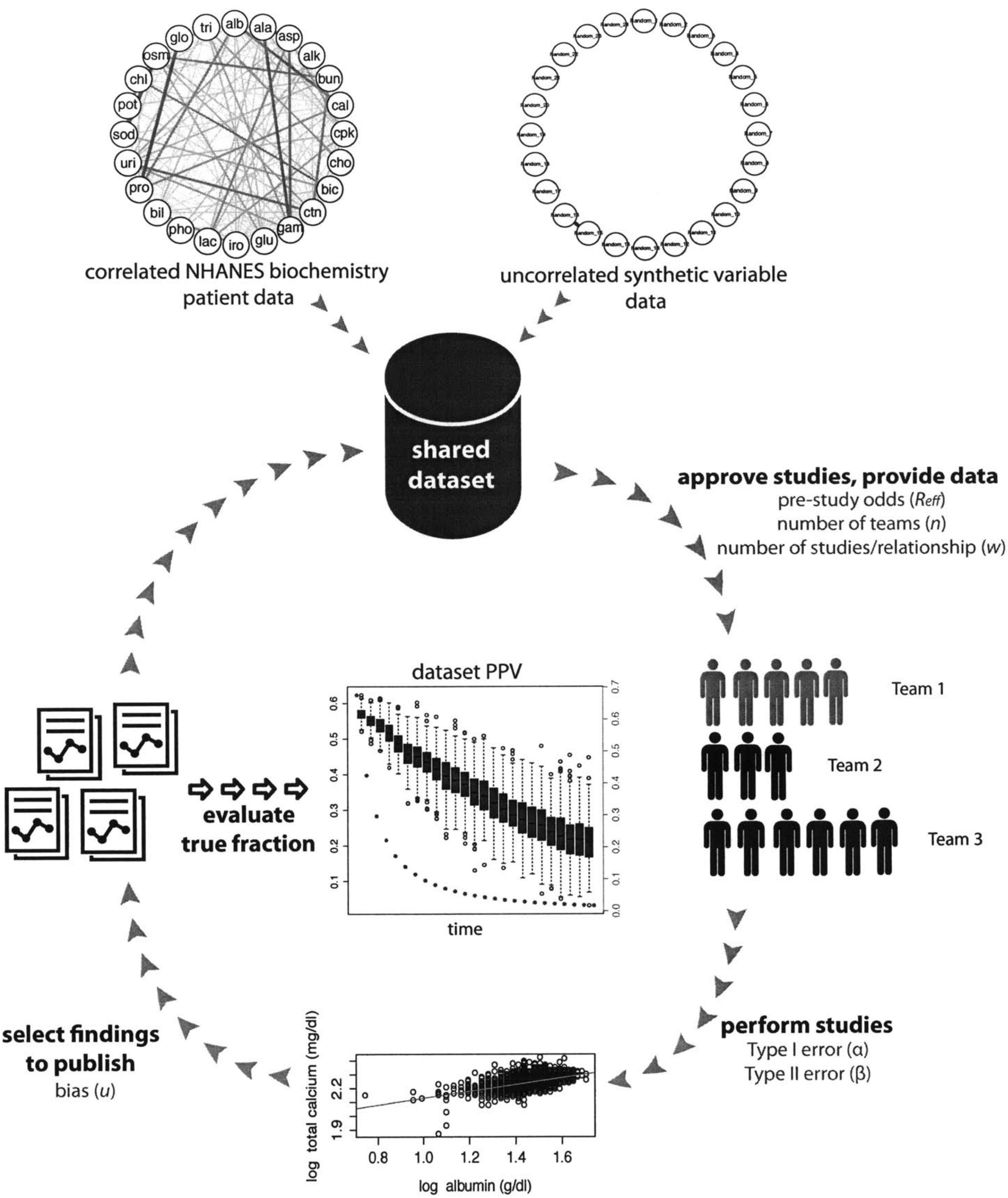
33. Gordon, D. et al. Publication of trials funded by the National Heart, Lung, and Blood Institute. *N. Engl. J. Med.* **369**, 1926–1934 (2013).

Table 1 Frequently encountered situations in communal science

Communal science situation	Example(s)	Affected *dPPV* parameter(s)
Late-arriving teams conduct studies with lower pre-study odds, power	dbGaP; secondary use of clinical trial data *Teams that access dbGaP datasets late may tend to study hypotheses without precedent (with smaller R_{eff}) or more complex hypotheses such as epistatic interactions (with larger β terms) (Figure 3a).*	R_{eff}, β
Intellectual crowding	iPSCs post Takahashi, Yamanaka 2006[32] *Many studies were spurred by the discovery of iPSCs, likely shifting N to greater "depth" (more studies per relationship). This impacts both the mean and the variance of dPPV, with the probability of a subsequent false discovery increasing with the degree of intellectual crowding (Figure 3c).*	**N**
Fluctuating funding	Political administration changes *Well-funded epochs of science, when many relationships are studied, have lower variance in dPPV. In poorly-funded times, with few relationships under study, dPPV has high variance (Figure 4b).*	**N**
Many smaller studies vs. fewer, larger studies	NIH Grants, NHLBI Clinical Trials between 2000-11,[33] NIH Large RFA vs. Investigator-Initiated Grant Programs *Funding agencies must balance the size and number of funded studies. β terms are smaller for large studies, but this likely comes at the expense of fewer relationships studied. Both mean and variance will be impacted in this tradeoff. Pilot studies may increase the mean of dPPV but may also inflate its variance (Figure 3b).*	**N, β**
Crowd-sourced contest	DREAM Challenges, Kaggle competitions *If many teams compete on a prediction task with no entry limit, the leaders' solutions may partly reflect overfitting, and performance will degrade on held out data (as seen in the Kaggle competitions: http://blog.kaggle.com/2012/07/06/the-dangers-of-overfitting-psychopathy-post-mortem/).*	*u*
Data analysts from different community than data generators	Pharmacovigilance, comparative effectiveness based on health claims data *Health claims data have established biases.[15] If such corrections are applied improperly as data are made available to analysts from different communities, u will increase, diminishing reproducibility (Figure 3c).*	*u*

Abbreviations: dbGaP: database of Genotypes and Phenotypes, iPSCs: induced pluripotent stem cells, NHLBI: National Heart, Lung, and Blood Institute, NIH: National Institutes of Health, RFA: Requests for Applications, DREAM: Dialogue on Reverse Engineering Assessment and Methods.

Framework for studying communal science



correlated NHANES biochemistry patient data

uncorrelated synthetic variable data

shared dataset

approve studies, provide data

pre-study odds (*Reff*)

number of teams (*n*)

number of studies/relationship (*w*)

perform studies

Type I error (α)

Type II error (β)

select findings to publish

bias (*u*)

evaluate true fraction

Figure 1: Framework for studying communal science. Correlation globes for NHANES 2011-12 Standard Biochemistry Profile measurements from ~6000 individuals and uncorrelated synthetic variables that were obtained by randomly sampling from the NHANES empirical distributions (line thickness is proportional to pairwise Pearson's r; only |r| > 0.05 are shown) . These data are combined to form a shared dataset. At each time step, multiple teams are "approved" to access the shared data, and then studies with Type I error α and Type II error β (which may be different for each study) and conducted. Findings are then "published" subject to bias *u*. The dataset positive predictive value, *dPPV*(t) is tracked during the simulations (center). Shown in red are relevant model parameters at each step.

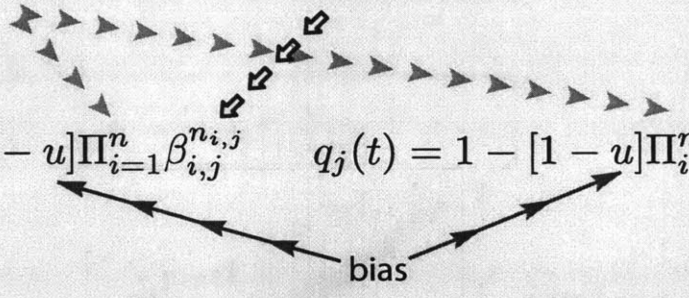
Box: The Dataset Positive Predictive Value (dPPV)

$$dPPV(t) = \frac{\mathrm{PB}(c_T(t), p_j(t))}{\mathrm{PB}(c_T(t), p_j(t)) + \mathrm{PB}(c_F(t), q_j(t))}$$

$$\mathrm{E}(dPPV(t)) \approx \frac{R_{\mathrm{eff}} \frac{1}{c_T} \sum_{j=1}^{c_T} 1 - [1-u]\Pi_{i=1}^{n} \beta_{i,j}^{n_{i,j}}}{R_{\mathrm{eff}} \frac{1}{c_T} \sum_{j=1}^{c_T} 1 - [1-u]\Pi_{i=1}^{n} \beta_{i,j}^{n_{i,j}} + \frac{1}{c_F} \sum_{j=1}^{c_F} 1 - [1-u]\Pi_{i=1}^{n} (1-\alpha)^{n_{i,j}}}$$

$$\mathrm{Var}[dPPV(t)] \approx \frac{\mu_{\mathrm{TP}}^2}{\mu_{\mathrm{TP+FP}}^2} \left(\frac{\mathrm{Var}(TP)}{\mu_{\mathrm{TP}}^2} + \frac{\mathrm{Var}(TP+FP)}{\mu_{\mathrm{TP+FP}}^2} - 2 \frac{\mathrm{Cov}(TP, TP+FP)}{\mu_{\mathrm{TP}} \mu_{\mathrm{TP+FP}}} \right)$$

$$\mathbf{N} = \begin{array}{c} c_m \text{ relationships} \\ \begin{bmatrix} n_{1,1} & n_{1,2} & \cdots & n_{1,c_m} \\ n_{2,1} & n_{2,2} & \cdots & n_{2,c_m} \\ \vdots & \vdots & \ddots & \vdots \\ n_{n,1} & n_{n,2} & \cdots & n_{n,c_m} \end{bmatrix} \end{array} \qquad \boldsymbol{\beta} = \begin{array}{cc} c_m \text{ relationships} & \\ \begin{bmatrix} \beta_{1,1} & \beta_{1,2} & \cdots & \beta_{1,c_m} \\ \beta_{2,1} & \beta_{2,2} & \cdots & \beta_{2,c_m} \\ \vdots & \vdots & \ddots & \vdots \\ \beta_{n,1} & \beta_{n,2} & \cdots & \beta_{n,c_m} \end{bmatrix} & \begin{array}{c} \text{n} \\ \text{teams} \end{array} \end{array}$$



$$p_j(t) = 1 - [1-u]\Pi_{i=1}^{n} \beta_{i,j}^{n_{i,j}} \qquad q_j(t) = 1 - [1-u]\Pi_{i=1}^{n} (1-\alpha)^{n_{i,j}}$$

The Dataset Positive Predictive Value *dPPV*(*t*) is a random variable representing the fraction of claimed relationships that are true at time *t*. *dPPV*(*t*) is a time-dependent ratio of two Poisson binomial distributions. At time *t*, there are $c_T(t)$ "non-null" relationships under study and $c_F(t)$ "null" relationships under study. The ratio of these two quantities is the pre-study odds, $R_{\mathrm{eff}}(t) = c_T(t)/c_F(t)$. The *n* x c_m binary matrix **N** specifies whether team *i* is studying relationship *j*, and the *n* x c_m matrix **β** gives the Type II errror of team *i*'s study of relationship *j*. These matrices are used to compute $p_j(t)$ and $q_j(t)$ at time *t*, which are vectors representing the relationship-wise Type II errors for the $c_T(t)$ "non-null" relationships and Type I errors for $c_F(t)$ "null" relationships under study, respectively. Relationships are claimed at significance level α. PB($c_T(t)$,$p_j(t)$) is a Poisson binomial random variable representing the number of true positives from $c_T(t)$ trials with probability vector $p_j(t)$, and PB($c_F(t)$,$q_j(t)$) is a Poisson binomial random variable representing the number of false positives from $c_F(t)$ trials with probability vector $q_j(t)$. Even with identical parameters, there may be high variability in the realized values of *dPPV*. This spread is captured by the variance of *dPPV*, whose approximation is provided above. The central tendency is captured by the expected value of *dPPV*.

Structural factors that influence the reproducibility of communal science

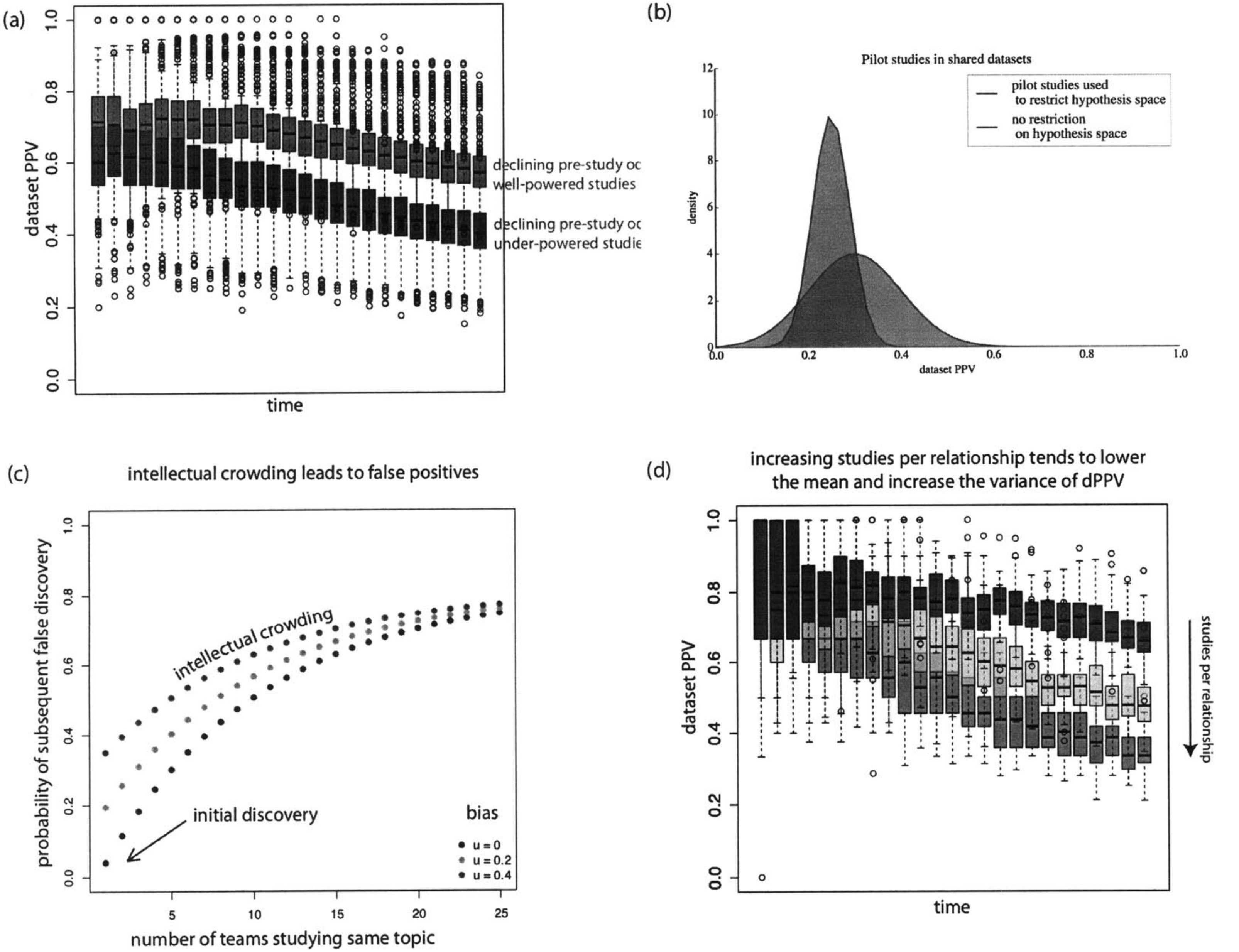


Figure 3: Structural factors that influence the reproducibility of communal science. Monte Carlo simulations using the NHANES data. **(a)** If late-arriving teams pursue hypotheses with lower pre-study odds, reproducibility (*dPPV*) tends to decline. Ensuring adequate power lessens the false discovery risks from communal pursuit of less likely hypotheses. **(b)** While pilot studies can improve the pre-study odds of future studies, they also tend to lead to greater variance in *dPPV*, placing greater density on very low reproducibility. **(c)** "Intellectual crowding": an initial discovery is followed by a wave of similar studies. In such situations the probability of a subsequent false discovery rises rapidly, with bias worsening the situation. **(d)** Increasing studies per relationships tends to lower the mean and increase the variance of *dPPV*.

Liberal data governance makes reproducibilty more stable

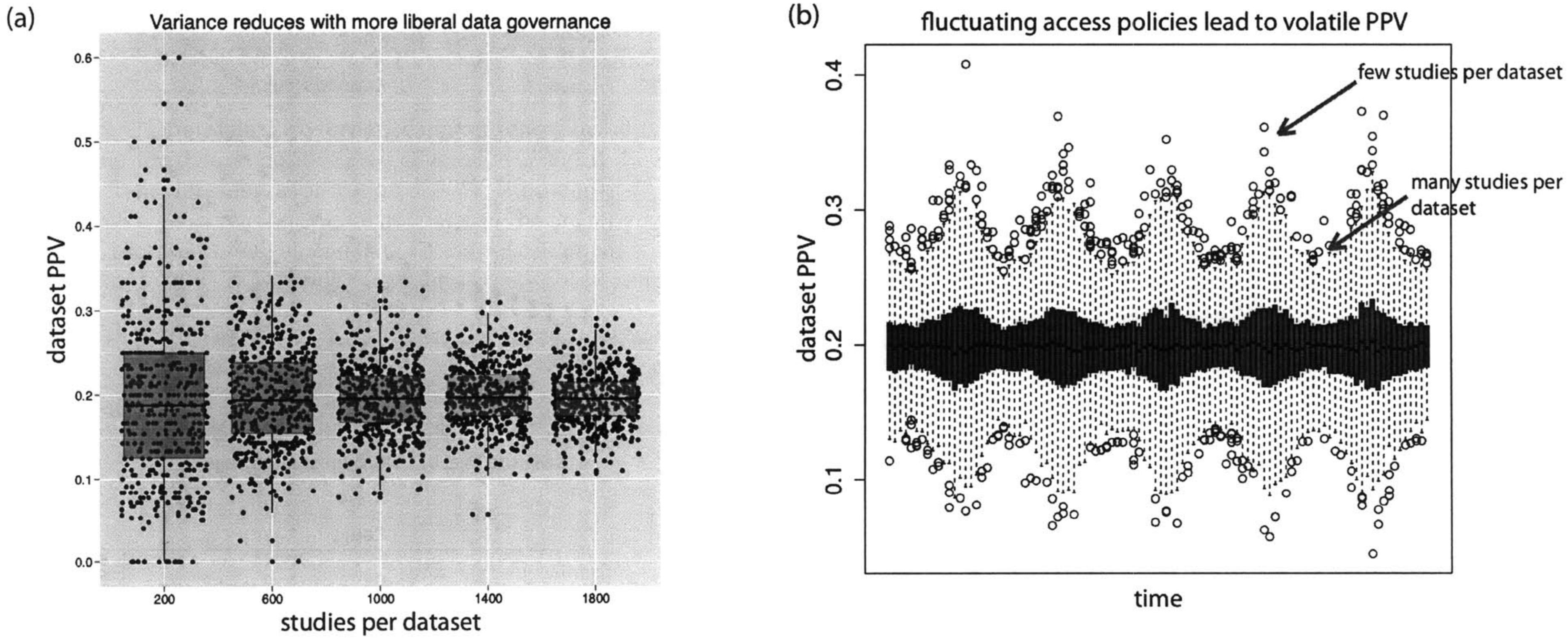


Figure 4: Liberal data governance makes reproducibility more stable. (a) The fraction of claims from communal use of a shared dataset that are true, the dataset positive predictive value (*dPPV*), is shown vs. the number of studies per dataset (simulated using the NHANES 2011-12 data). Each point represents a realization of *dPPV*. The mean of *dPPV* does not vary with the number of studies per dataset, but the variance of *dPPV* declines with increasing studies per dataset. The red represents more restrictive data access policies (few studies, teams per dataset) and the green represents more liberal data access policies (many studies, teams per dataset). **(b)** Fluctuating access policies (or interest in a field) lead to volatility in PPV. Equivalently, fluctuating access policies affect the *variance* of *dPPV* but have little effect on the *mean* of *dPPV*.

Chapter 4

The dataset positive predictive value ($dPPV$)

Overview

Here we briefly motivate and define $dPPV$, the dataset positive predictive value. We first review the model for PPV published by Ioannidis[1] and then introduce new parameters to generalize this model for the situation in which many teams are studying different relationships in a shared dataset with studies of differing power. We use these new parameters to define a new time-dependent equation for PPV. We then recast this version of PPV as the mean of a discrete-time stochastic process $dPPV$, the ratio of two correlated Poisson binomial random variables. We compute approximations of the mean and variance of $dPPV$ and describe Monte Carlo simulations with $dPPV$ using NHANES patient data.[2]

1 Review of the Ioannidis model (PLOS Medicine 2005)

The PPV model published by Ioannidis[1] uses the following notation:

- PPV: positive predictive value
- R: pre-study odds, the ratio of "non-null" relationships to "null" relationships among those tested
- α: Type I Error Rate
- β: Type II Error Rate
- u: bias
- n: number of teams
- c: number of relationships tested in the field

Using this notation, PPV in the absence of bias is given by:

$$PPV = \frac{(1-\beta)R}{R - \beta R + \alpha} \tag{1}$$

PPV with bias u:

$$PPV = \frac{(1-\beta)R + u\beta R}{R + \alpha - \beta R + u - u\alpha + u\beta R} \tag{2}$$

PPV with n teams:

$$PPV = \frac{R(1-\beta^n)}{R(1-\beta^n) + 1 - [1-\alpha]^n} \tag{3}$$

Model notes:

- Each study has equal power $(1-\beta)$ to detect any of the "non-null" relationships
- Parameters are fixed (e.g. n, R)

- For the many teams scenario, teams behave identically, each pursuing all c relationships

- For the many teams scenario, PPV is the fraction of claimed relationships that are true amongst all claimed relationships, where a relationship is "claimed" if *at least one* team claims it, and is not claimed only if *no team* claims it

- A large number of relationships are being tested such that the probabilities in the equations above may be considered *rates*, and we may neglect the variance in PPV (detailed below)

2 Introducing N and $\boldsymbol{\beta}$

In order to generalize the Ioannidis model[1] for the situation in which many teams may be studying different relationships with studies of differing power, we introduce two new matrix variables. Let **N** be an $n \times c_m$ matrix (m for "maximum" number of relationships) of binary variables where $n_{i,j}$ indicates whether team i is studying relationship j:

$$\mathbf{N} = \begin{array}{c} c_m \text{ relationships} \\ \begin{bmatrix} n_{1,1} & n_{1,2} & \cdots & n_{1,c_m} \\ n_{2,1} & n_{2,2} & \cdots & n_{2,c_m} \\ \vdots & \vdots & \ddots & \vdots \\ n_{n,1} & n_{n,2} & \cdots & n_{n,c_m} \end{bmatrix} \end{array} \begin{array}{c} \text{n} \\ \text{teams} \end{array} \tag{4}$$

Let $\boldsymbol{\beta}$ be an $n \times c_m$ matrix where $\beta_{i,j}$ specifies the Type II error of team i's study of relationship j. Equivalently, $1 - \beta_{i,j}$ is the power of team i's study of relationship j:

$$\boldsymbol{\beta} = \begin{array}{c} c_m \text{ relationships} \\ \begin{bmatrix} \beta_{1,1} & \beta_{1,2} & \cdots & \beta_{1,c_m} \\ \beta_{2,1} & \beta_{2,2} & \cdots & \beta_{2,c_m} \\ \vdots & \vdots & \ddots & \vdots \\ \beta_{n,1} & \beta_{n,2} & \cdots & \beta_{n,c_m} \end{bmatrix} \end{array} \begin{array}{c} \text{n} \\ \text{teams} \end{array} \tag{5}$$

where we set $\beta_{i,j} \equiv -1$ iff $n_{i,j} = 0$ (that is, iff the i^{th} group is not studying relationship j). Note that **N** and $\boldsymbol{\beta}$ are redundant ($n_{i,j} = 0$ iff $\beta_{i,j} = -1$ and $n_{i,j} = 1$ iff $\beta_{i,j} \neq -1$), but we keep the notation for clarity below.

The Hadamard (element-wise) product zeroes out the power terms for those studies not being conducted:

$$
\begin{aligned}
\mathbf{N} \odot \boldsymbol{\beta} &= \begin{bmatrix} n_{1,1} & n_{1,2} & \cdots & n_{1,c_m} \\ n_{2,1} & n_{2,2} & \cdots & n_{2,c_m} \\ \vdots & \vdots & \ddots & \vdots \\ n_{n,1} & n_{n,2} & \cdots & n_{n,c_m} \end{bmatrix} \odot \begin{bmatrix} \beta_{1,1} & \beta_{1,2} & \cdots & \beta_{1,c_m} \\ \beta_{2,1} & \beta_{2,2} & \cdots & \beta_{2,c_m} \\ \vdots & \vdots & \ddots & \vdots \\ \beta_{n,1} & \beta_{n,2} & \cdots & \beta_{n,c_m} \end{bmatrix} \\
&= \begin{bmatrix} n_{1,1}\beta_{1,1} & n_{1,2}\beta_{1,2} & \cdots & n_{1,c_m}\beta_{1,c_m} \\ n_{2,1}\beta_{2,1} & n_{2,2}\beta_{2,2} & \cdots & n_{2,c_m}\beta_{2,c_m} \\ \vdots & \vdots & \ddots & \vdots \\ n_{n,1}\beta_{n,1} & n_{n,2}\beta_{n,2} & \cdots & n_{n,c_m}\beta_{n,c_m} \end{bmatrix}
\end{aligned}
$$

3 Two types of PPV

For the many teams scenario,[1] Equation 3 can be written as:

$$
\begin{aligned}
PPV &= \frac{c\frac{R}{R+1}(1-\beta^n)}{c\frac{R}{R+1}(1-\beta^n) + c\frac{1}{R+1}(1-[1-\alpha]^n)} \\
&= \frac{R(1-\beta^n)}{R(1-\beta^n) + (1-[1-\alpha]^n)} \\
&= \frac{Rf(\beta)}{Rf(\beta) + f(\alpha)}
\end{aligned}
$$

where $f(\beta) = (1-\beta^n)$ and $f(\alpha) = 1-[1-\alpha]^n$. We consider the following method, analogous to Ioannidis[1] for defining $f(\alpha)$ and $f(\beta)$ in our model:

$$
\text{Model rates: } \begin{cases} f(\alpha) = (1/c_F)\sum_{c_F}(1-\Pi_{i=1}^{n}(1-\alpha)^{n_{i,j}}) \\ f(\beta) = (1/c_T)\sum_{c_T}(1-\Pi_{i=1}^{n}\beta_{i,j}^{n_{i,j}}) \end{cases} \tag{6}
$$

We could define $f(\alpha)$ and $f(\beta)$ to be rates of true and false *claims* made, several of which might be about the same relationship. Instead we follow Ioannidis[1] in defining $f(\alpha)$ and $f(\beta)$ as rates of true and false *relationships* claimed, though many similar results may be derived using either scheme. Note that in Equation 6 above, if $\beta_{i,j} = \beta$ everywhere then $f(\beta)$ simplifies to $(1 - \beta^n)$, in agreement with Ioannidis.[1]

4 Two types of data

The set of c relationships under study by n teams (specified by $\mathbf{N}$) might define a "field" where the data is separately sampled in each individuals team's study, or might be a shared dataset D. For most purposes, the second scenario is effectively a simplified version of the former, where β is identical for all teams pursuing the same relationship (assuming the inference procedures are equivalent). Moreover, PPV is simpler in this scenario—Method 1 reduces to:

$$\text{Model rates: } \begin{cases} f(\alpha) = \alpha \\ f(\beta) = (1/c_T) \sum\limits_{c_T} (1 - \beta_j) \end{cases} \tag{7}$$

where β_j is the power for studies of relationship j.

5 Introducing R_{eff} and θ

Let $R_{\text{eff}} \in [0, 1]$ be the pre-study odds of the relationships being studied by the n teams, as specified in $\mathbf{N}$. Note if the n teams are studying the c possible relationships at random and c is sufficiently large, $R_{\text{eff}} \approx R$. We can write:

$$PPV = \frac{R_{\text{eff}}(1/c_T) \sum_{c_T} (1 - \Pi_{i=1}^{n} \beta_{i,j}^{n_{i,j}})}{R_{\text{eff}}(1/c_T) \sum_{c_T} (1 - \Pi_{i=1}^{n} \beta_{i,j}^{n_{i,j}}) + (1/c_F) \sum_{c_F} (1 - \Pi_{i=1}^{n} (1 - \alpha)^{n_{i,j}})} \tag{8}$$

Let θ be a function that takes as input a research design (e.g. "exploratory", "novel", "confirmatory", "meta-analysis") and outputs a corresponding typical R_{eff} for that design.

6 Adding bias u

We adopt the definition of bias from Ioannidis:[1]

"First, let us define bias as the combination of various design, data, analysis, and presentation factors that tend to produce 'research findings' when they should not be produced. Let u be the proportion of probed analyses that would not have been 'research findings,' but nevertheless end up presented and reported as such, because of bias."[1]

Introducing u impacts the model rates as follows:

$$\text{Model rates:} \begin{cases} f(\alpha) = (1/c_F)\sum_{c_F}(1 - [1-u]\Pi_{i=1}^{n}(1-\alpha)^{n_{i,j}}) \\ f(\beta) = (1/c_T)\sum_{c_T}(1 - [1-u]\Pi_{i=1}^{n}\beta_{i,j}^{n_{i,j}}) \end{cases} \tag{9}$$

assuming u has no relationship with the $\beta_{i,j}$ terms or α.

7 Modeling time

We let $\mathbf{N}$ vary with time: $\mathbf{N}(t)$. As a result, R_{eff} also varies with time: $R_{\text{eff}}(t)$. There are many ways to define $\mathbf{N}(t)$. As an illustration, consider the following simple scenario: a new team begins to analyze the dataset at every time point and picks a relationship to study from the c_m that are available. Define $\theta(t)$ to be the research design used by team t. $\theta(t)$ specifies the pre-study odds.

$$\overset{\overset{\theta(t)}{\downarrow}}{\mathbf{N}(t)} = \overset{t=1}{\begin{bmatrix} n_{1,1} & n_{1,2} & \cdots & n_{1,c_m} \end{bmatrix}} \rightarrow \overset{t=2}{\begin{bmatrix} n_{1,1} & n_{1,2} & \cdots & n_{1,c_m} \\ n_{2,1} & n_{2,2} & \cdots & n_{2,c_m} \end{bmatrix}} \rightarrow \cdots \rightarrow \overset{t=t'}{\begin{bmatrix} n_{1,1} & n_{1,2} & \cdots & n_{1,c_m} \\ n_{2,1} & n_{2,2} & \cdots & n_{2,c_m} \\ \vdots & \vdots & \ddots & \vdots \\ n_{t',1} & n_{t',2} & \cdots & n_{t',c_m} \end{bmatrix}} \tag{10}$$

For example:

$$\overset{\substack{\theta(t)\\ \downarrow}}{\mathbf{N}(t)} = \overset{t=1}{\begin{bmatrix}0 & 1 & \cdots & 0\end{bmatrix}} \rightarrow \overset{t=2}{\begin{bmatrix}0 & 1 & \cdots & 0\\ 1 & 0 & \cdots & 0\end{bmatrix}} \rightarrow \cdots \rightarrow \overset{t=t'}{\begin{bmatrix}0 & 1 & \cdots & 0\\ 1 & 0 & \cdots & 0\\ \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & \cdots & 1\end{bmatrix}} \tag{11}$$

We also let $\boldsymbol{\beta}$ vary with time: $\boldsymbol{\beta}(t)$. Consider the following scenario: teams with initial access to the data pursue "simpler" relationships in the data (e.g. single-variable associations) while late-arriving teams pursue more "complex" relationships in the data (e.g. interactive effects). Controlling for Type I Error, the late-arriving investigators will tend to have reduced power to identify complex relationships. Alternatively, if late-arriving teams are able to recruit more individuals per study (perhaps by pooling data with another team or receiving additional funding for their study), later studies may be able to maintain power even as complexity increases.

$$\boldsymbol{\beta}(t) = \overset{t=1}{\begin{bmatrix}\beta_{1,1} & \beta_{1,2} & \cdots & \beta_{1,c_m}\end{bmatrix}} \rightarrow \overset{t=2}{\begin{bmatrix}\beta_{1,1} & \beta_{1,2} & \cdots & \beta_{1,c_m}\\ \beta_{2,1} & \beta_{2,2} & \cdots & \beta_{2,c_m}\end{bmatrix}} \rightarrow \cdots \rightarrow \overset{t=t'}{\begin{bmatrix}\beta_{1,1} & \beta_{1,2} & \cdots & \beta_{1,c_m}\\ \beta_{2,1} & \beta_{2,2} & \cdots & \beta_{2,c_m}\\ \vdots & \vdots & \ddots & \vdots\\ \beta_{t',1} & \beta_{t',2} & \cdots & \beta_{t',c_m}\end{bmatrix}} \tag{12}$$

For example, the following version of $\boldsymbol{\beta}(t)$ might correspond to the $\mathbf{N}(t)$ in Equation 11, where a new team begins analyzing the shared dataset at each time step and chooses a single relationship to study:

$$\boldsymbol{\beta}(t) = \overset{t=1}{\begin{bmatrix}-1 & 0.1 & \cdots & -1\end{bmatrix}} \rightarrow \overset{t=2}{\begin{bmatrix}-1 & 0.1 & \cdots & -1\\ 0.3 & -1 & \cdots & -1\end{bmatrix}} \rightarrow \cdots \rightarrow \overset{t=t'}{\begin{bmatrix}-1 & 0.1 & \cdots & -1\\ 0.05 & -1 & \cdots & -1\\ \vdots & \vdots & \ddots & \vdots\\ -1 & -1 & \cdots & 0.6\end{bmatrix}} \tag{13}$$

where -1 indicates no relationship is being pursued, mirroring zeroes in $\mathbf{N}$. In this example, as with $\mathbf{N}(t)$ in Equation 11, a row is added at every time step. $\beta_{i,j}(t)$ can change at any time point. For example, $\beta_{2,1}$ drops from 0.3 to 0.05 from $t = 2$ to $t = t'$, reflecting e.g. an increase in available sample size and corresponding gain in power.

8 PPV(t)

Tying everything together, we may now write a time-dependent version of PPV, which incorporates bias and the time dependence of $\boldsymbol{\beta}(t)$ and $R_{\text{eff}}(t)$:

$$\boxed{PPV(t) = \frac{\frac{R_{\text{eff}}(t)}{c_T(t)} \sum\limits_{c_T(t)} (1 - [1-u]\Pi_{i=1}^{n(t)} \beta_{i,j}(t)^{n_{i,j}})}{\frac{R_{\text{eff}}(t)}{c_T(t)} \sum\limits_{c_T(t)} (1 - [1-u]\Pi_{i=1}^{n(t)} \beta_{i,j}(t)^{n_{i,j}}) + \frac{1}{c_F(t)} \sum\limits_{c_F(t)} (1 - [1-u]\Pi_{i=1}^{n(t)} (1-\alpha)^{n_{i,j}})}} \tag{14}$$

We can show that:

$$\boxed{PPV(t) \text{ declines with decreasing } R_{\text{eff}}(t).}$$

Assuming $f(\alpha)$ and $f(\beta)$ unchanging, it is straightforward to show that iff $R_{\text{eff}}(t_1) \leq R_{\text{eff}}(t_2)$ for any two time points t_1 and t_2 then $PPV(t_1) \leq PPV(t_2)$. Note that Equation 14 has the form:

$$PPV(t) = \frac{R_{\text{eff}}(t) f(\beta)}{R_{\text{eff}}(t) f(\beta) + f(\alpha)} \tag{15}$$

where all terms are nonnegative. The equation $PPV(t_1) \leq PPV(t_2)$ can be written as:

$$\begin{aligned}
\frac{R_{\text{eff}}(t_1) f(\beta)}{R_{\text{eff}}(t_1) f(\beta) + f(\alpha)} &\leq \frac{R_{\text{eff}}(t_2) f(\beta)}{R_{\text{eff}}(t_2) f(\beta) + f(\alpha)} \\
R_{\text{eff}}(t_1) R_{\text{eff}}(t_2) f(\beta)^2 + R_{\text{eff}}(t_1) f(\alpha) f(\beta) &\leq R_{\text{eff}}(t_1) R_{\text{eff}}(t_2) f(\beta)^2 + R_{\text{eff}}(t_2) f(\alpha) f(\beta) \\
R_{\text{eff}}(t_1) f(\alpha) f(\beta) &\leq R_{\text{eff}}(t_2) f(\alpha) f(\beta) \iff R_{\text{eff}}(t_1) \leq R_{\text{eff}}(t_2)
\end{aligned}$$

given that all terms are nonnegative. □

Much as positive predictive value of a diagnostic test falls with declining pre-study odds, the PPV in a shared database falls as teams pursue *a priori* less likely hypotheses over time. This may happen as newly-arriving teams pursue novel or more complex hypotheses. This equation also sheds light on how we can improve $PPV(t)$: by increasing $f(\beta)$ or decreasing $f(\alpha)$, which generally correspond to reducing $\beta_{i,j}$ and α. $f(\beta)$ may be increased by increasing the power of each study or reducing the number of independent investigations of the same relationship, provided an analogy of the relationship $1-\beta<\alpha$ from[1] holds.

9 Data governance: $s(t)$, $c(t)$, and $n(t)$

The data governance policy is captured by three functions: $s(t)$, the number of studies at time t; $c(t)$, the number of relationships being studied at time t; and $n(t)$, the number of teams at time t. If $s(t)\approx c(t)$ then there is approximately one study per relationship. Some possible forms of $n(t)$ include:

$$n(t)=\begin{cases} a & \text{fixed number of teams} \\ t & \text{linearly increasing number of teams, as in Section 7} \\ at^{\lambda} & \text{decreasing number of teams if } \lambda<0\text{; increasing if } \lambda>0 \\ a\sin(bt) & \text{fluctuating access policies} \end{cases} \tag{16}$$

10 The dataset Positive Predictive Value, $dPPV(t)$

$PPV(t)$ in Equation 14 may be viewed as the *mean* of a discrete-time stochastic process $dPPV(t)$, the ***dataset* positive predictive value**:

$$dPPV(t)=\frac{TP(t)}{TP(t)+FP(t)} \tag{17}$$

where $TP(t)$ follows a Poisson binomial distribution with $c_T(t)=c(t)R_{\text{eff}}/(R_{\text{eff}}+1)$ trials and success probability vector $\vec{p}(t) \in [0,1]^{c_T(t)}$. Similarly, $FP(t)$ follows a Poisson binomial distribution with $c_F(t) = c(t)/(R_{\text{eff}} + 1)$ trials and success probability $\vec{q}(t) \in [0,1]^{c_F(t)}$. We can write $\vec{p}(t)$ as the vector of column-wise products of power terms from matrix $\boldsymbol{\beta}$: $\vec{p}(t) = p_j(t) = (1 - [1 - u]\Pi_{i=1}^{n} \beta_{i,1}^{n_{i,1}}, 1 - [1-u]\Pi_{i=1}^{n} \beta_{i,2}^{n_{i,2}}, \ldots, 1 - [1-u]\Pi_{i=1}^{n} \beta_{i,c_T}^{n_{i,c_T}})$ where we have replaced the vector notation with the subscript j that indexes the columns of $\boldsymbol{\beta}$. Similarly, $\vec{q}(t) = q_j(t) = (1 - [1 - u]\Pi_{i=1}^{n}(1 - \alpha)^{n_{i,1}}, 1 - [1 - u]\Pi_{i=1}^{n}(1 - \alpha)^{n_{i,2}}, \ldots, 1 - [1 - u]\Pi_{i=1}^{n}(1 - \alpha)^{n_{i,c_F}})$. More compactly:

$$dPPV(t) = \frac{\text{PB}(c_T(t), p_j(t))}{\text{PB}(c_T(t), p_j(t)) + \text{PB}(c_F(t), q_j(t))} \tag{18}$$

$$\begin{aligned} p_j(t) &= 1 - [1 - u]\Pi_{i=1}^{n} \beta_{i,j}^{n_{i,j}} && (19) \\ q_j(t) &= 1 - [1 - u]\Pi_{i=1}^{n}(1 - \alpha)^{n_{i,j}} && (20) \end{aligned}$$

where we have dropped the explicit time notation.

The mean and variance of a Poisson binomial distribution with n trials and success probabilities $\vec{p} = p_j$ is given by:

$$\begin{aligned} \mu &= \sum_{j=1}^{n} p_j && (21) \\ \sigma^2 &= \sum_{j=1}^{n} p_j(1 - p_j) && (22) \end{aligned}$$

Thus, the mean and variance of $TP(t)$ is given by:

$$\begin{aligned} \mu_{\text{TP}} &= \sum_{j=1}^{c_T} 1 - [1 - u]\Pi_{i=1}^{n} \beta_{i,j}^{n_{i,j}} && (23) \\ \sigma_{\text{TP}}^2 &= \sum_{j=1}^{c_T} (1 - [1 - u]\Pi_{i=1}^{n} \beta_{i,j}^{n_{i,j}})([1 - u]\Pi_{i=1}^{n} \beta_{i,j}^{n_{i,j}}) && (24) \end{aligned}$$

and similarly the mean and variance of $FP(t)$ is given by:

$$\mu_{\text{FP}} = \sum_{j=1}^{c_F} 1 - [1-u]\Pi_{i=1}^{n}(1-\alpha)^{n_{i,j}} \tag{25}$$

$$\sigma^2_{\text{FP}} = \sum_{j=1}^{c_F}(1 - [1-u]\Pi_{i=1}^{n}(1-\alpha)^{n_{i,j}})([1-u]\Pi_{i=1}^{n}(1-\alpha)^{n_{i,j}}) \tag{26}$$

We note that $TP(t) + FP(t)$ is also Poisson Binomial with $c(t)$ trials and probability vector $(p_1, p_2, \dots, p_{c_T}, q_1, q_2, \dots, q_{c_F})$ ordered as "non-null" followed by "null" relationships.

Then $dPPV$ is the ratio of two (correlated) Poisson binomial random variables. We can write an instructive approximation using the delta method and Taylor expansions. Simulations have been conducted to validate these approximations. It can be shown that the expected value of $dPPV$ is approximately:

$$\text{E}(dPPV) \approx \frac{\mu_{\text{TP}}}{\mu_{\text{TP}} + \mu_{\text{FP}}} \tag{27}$$

where we note that μ_{TP} and μ_{FP} are functions of time depending on which studies are being conducted and with what power (Equations 23 and 25).

Note that we may rewrite Equation 27 as:

$$\begin{aligned}
\text{E}(dPPV) &\approx \frac{\mu_{\text{TP}}}{\mu_{\text{TP}} + \mu_{\text{FP}}} \\
&= \frac{\frac{R_{\text{eff}}c}{R_{\text{eff}}+1}\frac{1}{c_T}\mu_{\text{TP}}}{\frac{R_{\text{eff}}c}{R_{\text{eff}}+1}\frac{1}{c_T}\mu_{\text{TP}} + \frac{c}{R_{\text{eff}}+1}\frac{1}{c_F}\mu_{\text{FP}}} \\
&= \frac{R_{\text{eff}}\frac{1}{c_T}\mu_{\text{TP}}}{R_{\text{eff}}\frac{1}{c_T}\mu_{\text{TP}} + \frac{1}{c_F}\mu_{\text{FP}}}
\end{aligned}$$

where all terms vary with time (explicit notation omitted). Expanding this equation shows its equivalence to Equation 14 derived earlier, strengthening the framework for $dPPV$ as a stochastic process:

$$\mathrm{E}(dPPV(t)) \approx \frac{R_{\text{eff}}\frac{1}{c_T}\sum_{j=1}^{c_T} 1-[1-u]\Pi_{i=1}^{n}\beta_{i,j}^{n_{i,j}}}{R_{\text{eff}}\frac{1}{c_T}\sum_{j=1}^{c_T} 1-[1-u]\Pi_{i=1}^{n}\beta_{i,j}^{n_{i,j}} + \frac{1}{c_F}\sum_{j=1}^{c_F} 1-[1-u]\Pi_{i=1}^{n}(1-\alpha)^{n_{i,j}}} \tag{28}$$

In order to study the effect of governance policy, let us now hold R_{eff} constant. We also hold the "average power" γ_{avg} constant initially:

$$\gamma_{\text{avg}} \equiv \mu_{\text{TP}}/c_T = 1/c_T\sum_{j=1}^{c_T} 1-[1-u]\Pi_{i=1}^{n}\beta_{i,j}^{n_{i,j}} \tag{29}$$

10.1 Increasing power offsets reducing pre-study odds

Note that if R_{eff} and γ_{avg} are inversely related to one another in Equation 27. Assuming that the rate of false discoveries $\mu_{\text{FP}}/c_{\text{F}}$ is unchanging, then doubling the *relationship-averaged power* offsets the pursuit of hypotheses that are *a priori* half as likely.

10.2 Variance vs. number of relationships being studied

We can show that, in general:

$$\boxed{\text{The variance of } PPV(t) \text{ is inversely related to } c(t) = c_T(t) + c_F(t)}$$

We use a Taylor expansion and the delta method to approximate the variance of $dPPV(t)$. We find:

$$\mathrm{Var}[dPPV(t)] \approx \frac{\mu_{\text{TP}}^2}{\mu_{\text{TP+FP}}^2}\left(\frac{\mathrm{Var}(TP)}{\mu_{\text{TP}}^2} + \frac{\mathrm{Var}(TP+FP)}{\mu_{\text{TP+FP}}^2} - 2\frac{\mathrm{Cov}(TP, TP+FP)}{\mu_{\text{TP}}\mu_{\text{TP+FP}}}\right) \tag{30}$$

If $(1/c_T)\mu_{\text{TP}}$ and $(1/c_F)\mu_{\text{FP}}$ remain constant with time, and pre-study odds is not changing, then the variance of $dPPV(t)$ is inversely related to $c(t)$ even as the mean of $dPPV$ (PPV) is largely unaffected.

11 Correlation analyses and patient data

Here we give a concrete application of this framework by computing correlations between patient laboratory measurements. We use the NHANES 2011-12 Standard Biochemistry Profile data for these simulations (http://wwwn.cdc.gov/nchs/nhanes/). We use the following notation:

N	number of individuals
p	number of measurements per individual
r	sample correlation
D	complete data matrix, is $N \times p$ unless otherwise specified

11.1 Finding statistically significant correlations in random data

We use Pearson's r as our test statistic:

$$\boxed{r_{xy} = \frac{\mathrm{Cov}(x,y)}{s_x s_y} = r_{yx} = r} \tag{31}$$

If the data arise from an uncorrelated bivariate normal distribution, then the following variable t follows Student's t-distribution with $(n-2)$ degrees of freedom.

$$\boxed{t = r\sqrt{\frac{n-2}{1-r^2}}} \tag{32}$$

Inverting this statistic, we find that the critical value r to reach statistical significance depends on the desired level of significance indicated by t and the sample size n according to:

$$\boxed{r = \frac{t}{\sqrt{n-2+t^2}}} \tag{33}$$

For example, for $n = 10$ and $n = 1000$ which correspond to Student's t-distributions with critical values of $t_{df=8} = 2.228$ and $t_{df=998} = 1.962$, we have critical values of of $\boxed{r = 0.619}$ and $\boxed{r = 0.0620}$, respectively.

11.2 Simulating patient data

Given a positive-definite covariance matrix Σ, we may simulate patient data for N patients with p measurements according to a multivariate normal:

$$\boxed{D \sim \frac{1}{(2\pi)^{k/2}|(\Sigma)|^{1/2}} \exp\{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\}} \tag{34}$$

11.3 Decreasing $PPV(t)$ with decreasing $R_{\mathbf{eff}}(t)$

Below are the key steps in simulating using the NHANES data and $\mathbf{N}(t)$ to run a Monte Carlo simulation for teams arriving at the dataset sequentially. In this example, teams gaining access later to the dataset, perhaps under pressure of novelty, search over an increasingly *a priori* unlikely hypothesis space.

```
Data: $N \times p$ matrix of patient data
initialization;
while new team do
    team computes correlation for their study, late-arriving teams look in subsets of data
    with lower pre-study odds;
    if $p < 0.05$ then
        claim relationship;
    else
        do not claim relationship;
    end
end
using known correlation matrix ("gold standard"), look up true positives and false positives;
```

Algorithm 1: Monte Carlo simulation for long-term publication patterns when late-arriving teams study *a priori* less likely hypotheses

11.4 Increasing variance in $dPPV(t)$ with decreasing $s(t)$

```
Data: $N \times p$ matrix of patient data; Data governance policies for $I$ institutions
initialization;
while new institution do
    while new team do
        team computes correlation for their study;
        if $p < 0.05$ then
            claim relationship;
        else
            do not claim relationship;
        end
    end
end
look up and collate results;
```

Algorithm 2: Monte Carlo simulation for many institutions with different governance policies

References

[1] John P A Ioannidis. Why most published research findings are false. *PLoS medicine*, 2(8):e124, August 2005.

[2] National Center for Health Statistics (NCHS). National Health and Nutrition Examination Survey Questionnaire (or Examination Protocol, or Laboratory Protocol). (Centers for Disease Control and Prevention (CDC)) at http://www.cdc.gov/nchs/nhanes/

Chapter 5

Medicine's Uncomfortable Relationship with Math: Calculating Positive Predictive Value

Overview

In 1978, Casscells et. al.[1] published a small but important study showing that the majority of physicians, house officers, and students overestimated the positive predictive value of a laboratory test result using prevalence and false positive rate. Today, interpretation of diagnostic tests is even more critical with the increasing use of medical technology in healthcare. Accordingly, we replicated the Casscells study by asking a convenience sample of physicians, house officers, and students the same question: "If a test to detect a disease whose prevalence is 1/1000 has a false positive rate of 5 percent, what is the chance that a person found to have a positive result actually has the disease, assuming you know nothing about the person's symptoms or signs?"

METHODS

During July 2013, we surveyed a convenience sample of 24 attending physicians, 26 house officers, 10 medical students, and 1 retired physician at a Boston-area hospital, across a wide range of clinical specialties. Assuming a perfectly sensitive test, the correct answer is 1.96%; we considered "2%", "1.96%", or "<2%" correct. P-values were computed using the two-sided exact binomial test.

RESULTS

Approximately three-quarters of respondents answered the question incorrectly ($p < 0.001$). In our study, 14 of 61 respondents (23%) gave a correct response, not significantly different from the 11 of 60 (18%) correct responses in the Casscells study ($p = 0.32$). "95%" was the most common answer in both studies, given by 27 of 61 respondents (44%) in our study and 27 of 60 (45%) in Casscells et al. (Figure 1). We obtained a range of answers from "0.005%" to "96%", with a median of 66%, which is 33 times larger than the true answer. In brief explanations of their answers, respondents often knew to compute PPV but accounted for prevalence incorrectly. "PPV does not depend on prevalence", wrote one attending cardiologist, while, quite remarkably, a resident expected "better PPV when prevalence is low".

DISCUSSION

With wider availability of medical technology and diagnostic testing, sound clinical management will increasingly depend on statistical skills. We measured a key facet of statistical reasoning in practicing physicians and trainees: the evaluation of positive predictive value (PPV). Understanding PPV is particularly important when screening for unlikely conditions, where even nominally sensitive and specific tests can be diagnostically uninformative. Our results show that the majority of respondents in this single-hospital study could not assess PPV in the described scenario. Moreover, the most common error was a large overestimation of PPV, an error that could heavily impact the course of diagnosis and treatment.

We advocate increased training on evaluating diagnostics in general. Statistical reasoning was recognized to be an important clinical skill over 35 years ago[1-3] and notable initiatives like an HHMI-AAMC collaboration have developed recommendations to improve the next generation of medical education.[4,5] Our results suggest that these efforts, while laudable, could benefit from increased focus on statistical inference. Specifically, we favor revising premedical education standards to incorporate training in statistics in favor of calculus, which is seldom used in clinical practice. In addition, the practical applicability of medical statistics should be demonstrated throughout the continuum of medical training—not just in medical school.

To make use of these skills, clinicians need access to accurate sensitivity and specificity measures for ordered tests. In addition, we support software integrated into the electronic ordering system that can prevent common errors, and point-of-care resources like smartphones that can aid in calculation and test interpretation. The increasing diversity of diagnostic options promises to empower physicians to improve care if medical education can deliver the statistical skills needed to accurately incorporate these options into clinical care.

REFERENCES

1. Casscells W, Schoenberger A, Graboys TB. Interpretation by physicians of clinical laboratory results. *N. Engl. J. Med.* 1978;299(18):999–1001.

2. Berwick DM, Fineberg HV, Weinstein MC. When doctors meet numbers. *The American Journal of Medicine*. 1981;71(6):991–998.

3. Elstein AS. Heuristics and biases: selected errors in clinical reasoning. *Acad Med.* 1999;74(7):791–794.

4. *MR5: 5th Comprehensive Review of the Medical College Admission Test® (MCAT®).* Association of American Medical Colleges Available at: https://www.aamc.org/initiatives/mr5/. Accessed October 9, 2013.

5. Association of American Medical Colleges, Howard Hughes Medical Institute. Scientific Foundations for Future Physicians. 2009:1–46.

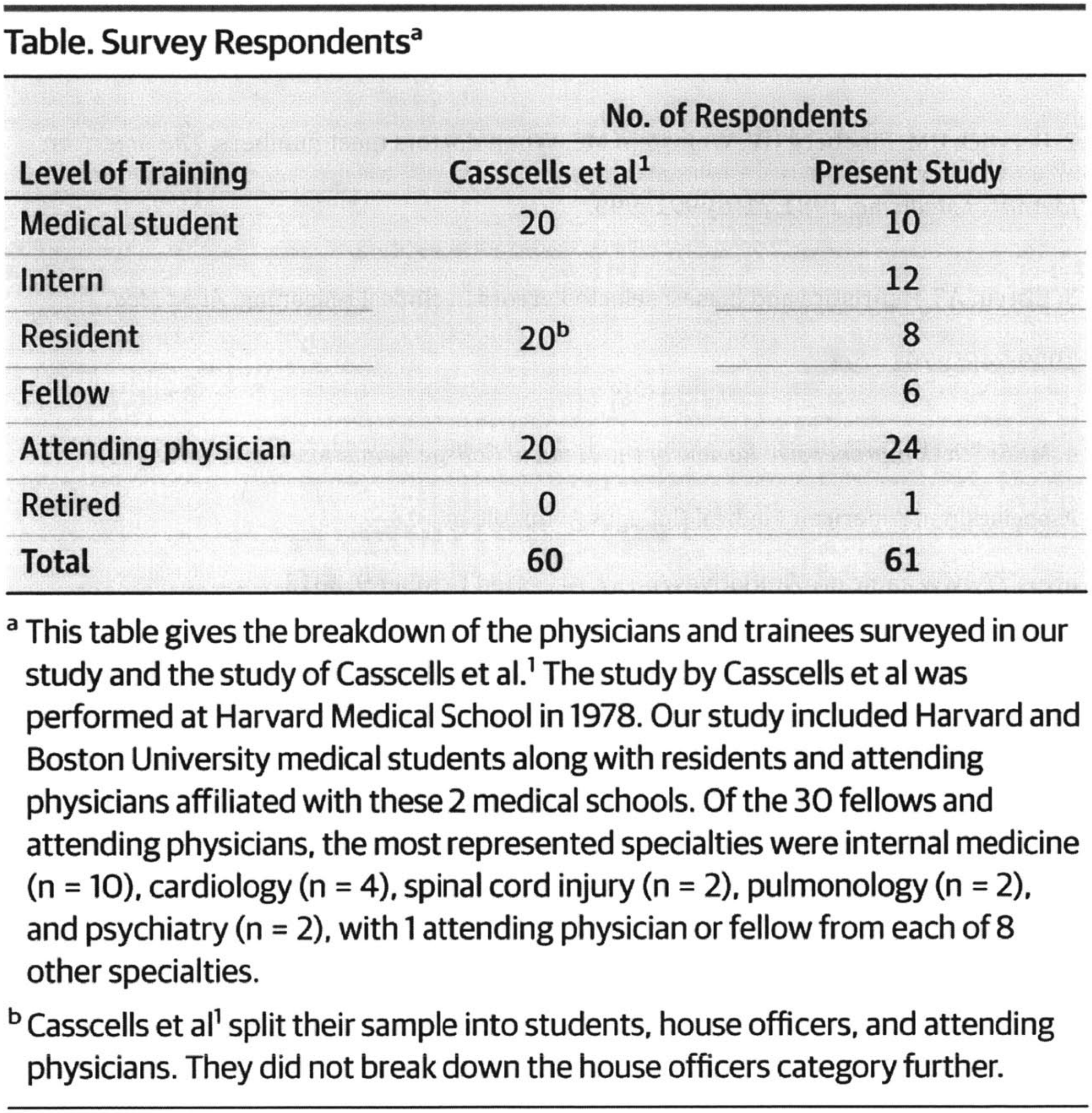


Table. Survey Respondents[a]

	No. of Respondents	
Level of Training	Casscells et al[1]	Present Study
Medical student	20	10
Intern	20[b]	12
Resident		8
Fellow		6
Attending physician	20	24
Retired	0	1
Total	**60**	**61**

[a] This table gives the breakdown of the physicians and trainees surveyed in our study and the study of Casscells et al.[1] The study by Casscells et al was performed at Harvard Medical School in 1978. Our study included Harvard and Boston University medical students along with residents and attending physicians affiliated with these 2 medical schools. Of the 30 fellows and attending physicians, the most represented specialties were internal medicine (n = 10), cardiology (n = 4), spinal cord injury (n = 2), pulmonology (n = 2), and psychiatry (n = 2), with 1 attending physician or fellow from each of 8 other specialties.

[b] Casscells et al[1] split their sample into students, house officers, and attending physicians. They did not break down the house officers category further.

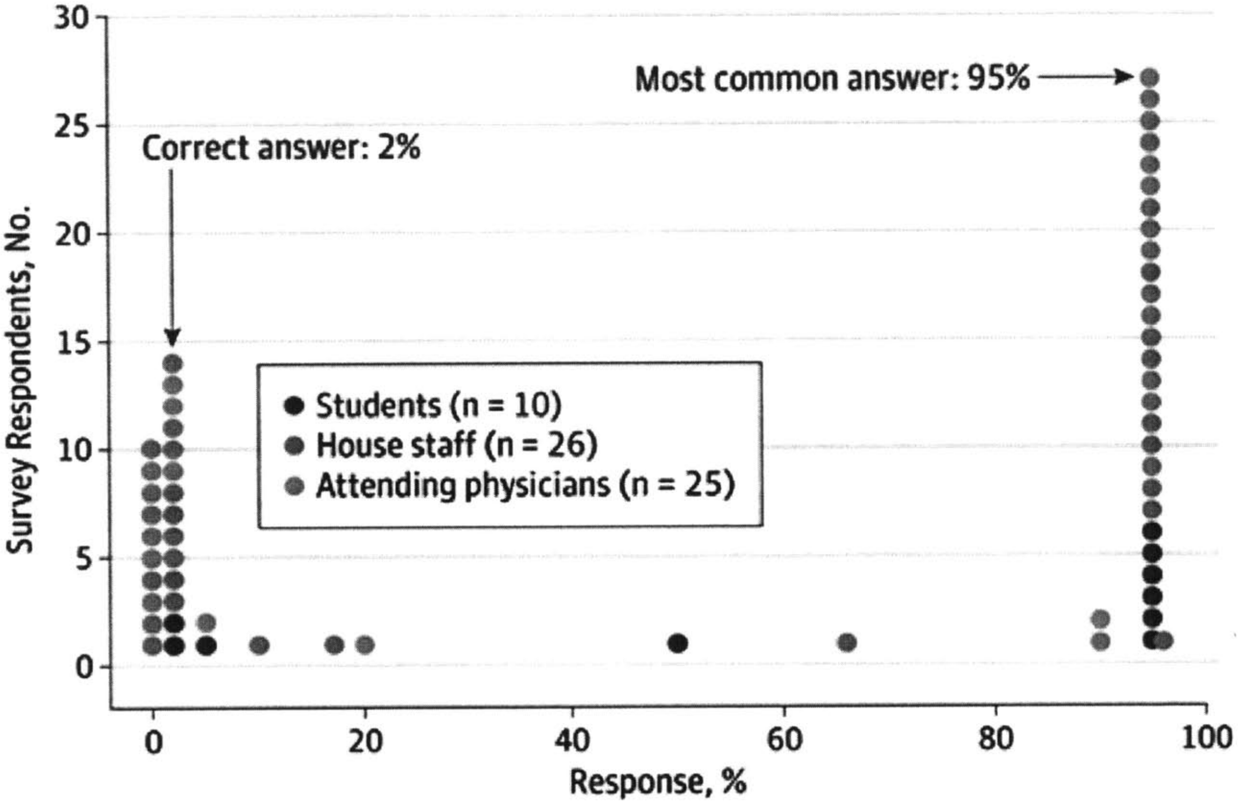


Figure.
Distribution of Responses to Survey Question Provided in the Article Text

Of 61 respondents, 14 provided the correct answer of 2%. The most common answer was 95%, provided by 27 of 61 respondents. The median answer was 66%, which is 33 times larger than the true answer.

This page is intentionally left blank.

Chapter 6: Conclusions and Future Directions

Towards a decision theoretic genomic medicine

In this thesis, I explored several statistical opportunities and challenges central to clinical decision-making and knowledge advancement using shared "big data" omics resources. In Chapter 2, I studied reclassifications of genetic variant pathogenicity ratings in hypertrophic cardiomyopathy, and showed how African Americans were disproportionately affected by false positive genomic variant misclassifications, demonstrating the need for diverse control data to vet variants going forward. In Chapters 3 and 4, I studied the rising practice of sharing genomic and other big data resources, a practice that mirrors genomic testing in scale but not synchrony. In order to quantify reproducibility in this context, I introduced a new quantity called the "dataset positive predictive value" (*dPPV*). Most recent efforts to enhance reproducibility emphasize the responsibility of individual researchers;[1,2] however, *dPPV* shows that there are structural factors such as the data governance policy that can be equally influential. Lastly, in Chapter 5, I surveyed a group of practicing physicians to assess statistical literacy, a skill that will become increasingly relevant as medicine incorporates the burgeoning catalogue of precision diagnostics and therapeutics into practice.[3]

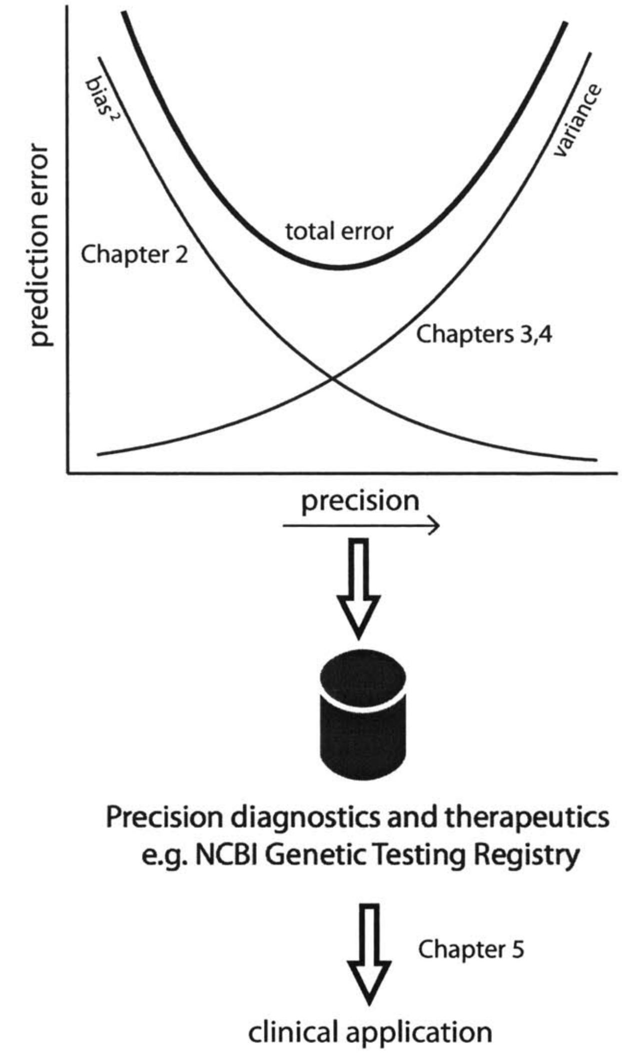


Figure 1: Mapping of thesis chapters onto the bias-variance tradeoff.

The findings in this thesis may be viewed as examples of a broader statistical tradeoff between bias and variance, a tension which pervades all facets of precision medicine. Recall that the error in future prediction of any statistical model can be decomposed into the sum of bias (squared) plus variance (plus irreproducible error),[4] and that increasing model complexity reduces bias at the expense of variance. In the context of an Information Commons,[5] as we correlate the high-dimensional patient state vector with disease, we must be wary of this inherent tradeoff—we must not be more precise than the data will allow us to be. Indeed the chapters of this thesis may be mapped onto this axis (Figure 1). Chapter 2 examines the adverse consequences of *bias*ed inference when findings derived from the study of one ethnic group are invalidated in another. In Chapters 3 and 4, I computed both the expectation and the *variance* of the dataset positive predictive value (*dPPV*) to study the central tendency and stability, respectively, of the true fraction of findings discovered by a community of researchers mining a shared dataset. Finally, Chapter 5 assesses a skill necessary for practitioners to translate research findings into improvements in clinical care.

The findings in this thesis suggest tractable ways to study and enhance reproducibility when researchers mine the Information Commons,[5] but equally important are methods to integrate the resulting precision diagnostics[3] into clinical care. I believe the most promising approach going forward will be found by looking to past, particularly at successful applications of *decision analysis* in medicine.[6-8] Decision analysis offers a principled approach to integrate the concepts of probability and utility, and perhaps most important for contemporary genomic medicine, it removes confusion during disagreement. In the words of Stephen Pauker and Jerome Kassirer,[6]

> Decision analysis is explicit; it forces us to consider all pertinent outcomes, it lays open in stark fashion all our assumptions about a clinical problem, including numerical representations of the chances and values of outcomes; it forces us to consider how patients feel about the quality of outcomes; and it allows us to come to grips precisely with the reasons why colleagues differ about actions to be taken. (Pauker, Kassirer 1987)

Developing a decision analytic framework on top of the Information Commons will foster reproducible use of the growing amounts of molecular, environmental, and clinical data available today. Achieving these goals will allow patients, researchers, and physicians alike to realize the full promise of the data-centric view of precision medicine.

REFERENCES

1. Collins, F.S. & Tabak, L.A. Policy: NIH plans to enhance reproducibility. *Nature* **505**, 612–613 (2014).

2. McNutt, M. Reproducibility. *Science* **343**, 229–229 (2014).

3. Rubinstein, W.S. et al. The NIH genetic testing registry: a new, centralized database of genetic tests to enable access to comprehensive information and improve transparency. *Nucleic Acids Research* **41**, D925–D935 (2012).

4. Hastie, T.J., Tibshirani, R.J. & Friedman, J.J.H. *The Elements of Statistical Learning*. 745 (Springer: 2009).

5. *Toward Precision Medicine: Building a Knowledge Network for Biomedical Research and a New Taxonomy of Disease.* 121 (National Academies Press: 2011).

6. Pauker, S.G. & Kassirer, J.P. Decision analysis. *New England Journal of Medicine* **316**, 250–258 (1987).

7. Pauker, S.G. & Kassirer, J.P. The Threshold Approach to Clinical Decision Making. *New England Journal of Medicine* **302**, 1109–1117 (1980).

8. Szolovits, P. & Pauker, S.G. Categorical and probabilistic reasoning in medical diagnosis. *Artificial Intelligence* **11**, 115–144 (1978).