ABSTRACT

Title of Dissertation:	WHO, WHAT, WHEN, WHERE, AND WHY? QUANTIFYING AND UNDERSTANDING BIOMEDICAL DATA REUSE
	Lisa Federer, Doctor of Philosophy, 2019
Dissertation directed by:	Dr. Katie Shilton, Associate Professor & Doctoral Program Director, College of Information Studies

Since the mid-2000s, new data sharing mandates have led to an increase in the amount of research data available for reuse. Reuse of data benefits the scientific community and the public by potentially speeding scientific discovery and increasing the return on investment of publicly funded research. However, despite the potential benefits of reuse and the increasing availability of data, research on the impact of data reuse is so far sparse. This dissertation provides a deeper understanding of the impacts of shared biomedical research data by exploring who is reusing data and for what purpose. Specifically, this dissertation examines use requests and dataset descriptions from three biomedical repositories that require potential requestors to submit descriptions of their planned reuse. Content analysis of use requests yields insight into who is requesting data and the methods and topics of their planned reuse. Comparing use requests to the descriptions of the original datasets provides insight into the breadth of impact of data reuse and text mining of the original dataset descriptions helps determine the topics of datasets that are highly reused. This study demonstrates that patterns of reuse differ between dataset types, with genomic

datasets used more frequently together in meta-analyses for topics that diverge from the original purpose of collection, while clinical datasets are used more often on their own within a context that is similar to the reason for which they were collected. While requestors do come from a range of career stages from around the world, they are not evenly distributed; most requests come from English-speaking countries, especially the United States. This study also finds that datasets that receive the most requests soon after release continue to go on to be more requested, and that datasets covering common diseases are requested more than datasets on rare diseases. These findings have implications for several stakeholders, including funders and institutions developing policies to reward and incentivize data sharing, researchers who share data and those who reuse it, and repositories and data curators who must make choices about which datasets to curate and preserve.

WHO, WHAT, WHEN, WHERE, AND WHY? QUANTIFYING AND UNDERSTANDING BIOMEDICAL DATA REUSE

by

Lisa Federer

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park, in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2019

Advisory Committee:
Professor Katie Shilton, Chair
Professor Lindley Darden, Dean's Representative
Professor Beth St. Jean
Professor Yla Tausczik
Professor Susan Winter

Acknowledgements

This dissertation is the culmination of intense research conducted over a very short period of time, and it was only possible with the help and support of a number of people whom I gratefully acknowledge here.

First, I thank my advisor, Dr. Katie Shilton, who has been tremendously supportive throughout this process. Some advisors might have tried to talk me out of my very ambitious dissertation timeline, but she was always encouraging and had incredibly helpful advice. I am likewise grateful for my committee members, Dr. Lindley Darden, Dr. Beth St. Jean, Dr. Yla Tausczik, and Dr. Susan Winter, who have given excellent feedback that has helped shape this dissertation and who were generous with their time in working with my tight schedule. I also thank Dr. Andrea Wiggins, who was my advisor when I started in the doctoral program but left for a new position after my first year. The work I did with her during that time was foundational to the research presented here, and I greatly appreciate her insights and direction.

I am very grateful to Dr. Mike Feolo from dbGaP, Sean Coady from NHLBI's BioLINCC repository, and Sharon Lawlor the NIDDK Central Repository, who were crucial in assisting me with obtaining the use requests and repository information that were the foundation of this research. I also appreciate Jim Mork's guidance in preparing my data for batch processing by the Medical Text Indexer, and I thank Dr.

Maryam Zaringhalam and Franklin Sayre for serving as additional coders to validate my qualitative coding methodology.

Undertaking a doctorate degree is an arduous task on its own; doing so while also employed in a demanding full-time position presents an additional set of challenges. I am fortunate to have had the enthusiastic support of leadership in both of the libraries where I have held positions while working on this degree. Dr. Keith Cogdill encouraged me when I first began considering pursuing the degree, and I appreciate his support throughout the process, as well as the support of all my former colleagues at the NIH Library. Since moving to the National Library of Medicine, I have been grateful for the encouragement and mentorship of Dr. Mike Huerta and Dr. Patti Brennan, as well as of the many colleagues at the NLM and the NIH who have provided helpful input and cheered me on throughout the process, especially Dianne Babski and Anna Ripple. I am also immensely appreciative of the financial support from the NIH Library and NLM that have made it possible for me to complete this degree.

I gratefully acknowledge the many friends and colleagues who have provided advice and feedback on my dissertation research, particularly Dr. Ben Busby, Dr. Maryam Zaringhalam, and the members of the Ethics and Values in Design lab. I also thank my non-academic friends who have been willing to listen to me ramble at length about data and have been enthusiastic cheerleaders all along the way, especially Ali Sabzevari, Susie Nguyen, Monica Waterston, Matt Woodrum, and Rich McGowan.

I thank my family for their unending support throughout not only this degree, but my many educational endeavors. My parents have always made clear their pride and love for me, and I am grateful to them for encouraging me in my academic interest from a young age. My father passed away during the first semester of my doctoral studies, and his absence at my graduation will be deeply felt, but I know he would have been very proud. I have been very thankful for the love and friendship of my mother during this degree and throughout my life.

Finally, even though I know she'll never read this, I thank my dog Ophelia, my best friend and faithful companion for the last seven years. She provided many cuddles and listened very seriously when I talked through my research problems with her. Most importantly, she never hesitated to remind me that, no matter how much serious work you have to do, you should always be sure you make time to play.

Table of Contents

List of Figures

List of Tables

Chapter 1: Introduction

In 2007, computer scientist Jim Gray asserted that the practice of science had been fundamentally changed by the advent of new technologies that facilitated the collection, storage, and analysis of large digital datasets. "Techniques and technologies for such data-intensive science are so different," he argued, "that it is worth distinguishing data-intensive science…as a new, fourth paradigm for scientific exploration" (Hey, Tansley, & Tolle, 2009, p. xix). In the decade since Gray first proposed this new paradigm, thousands of human genomes have been sequenced, and petabytes' worth of scientific data collected, with more pouring in every day, giving rise to a veritable data deluge.

It is not only the technical ability to more quickly and inexpensively gather, create, and store data that has transformed the practice of science, but also the establishment by both major funders and prominent publishing groups of mandates to share those data. Researchers around the world have begun to share their data not only in response to such mandates, but also as part of a growing movement toward open science practices that bring not only data, but a broad range of products of scientific research out of desk drawers and hard drives and into the public sphere, where they can be accessed, reused, and repurposed. In many fields, researchers today can feasibly conduct studies using publicly shared data, without ever having to set foot into a lab or seek funding to gather new data.

Despite this increasing availability of a broad range of datasets across scientific disciplines, little research has focused on how, why, or even *if* researchers are utilizing publicly available, shared research data. This dissertation aims to help close that gap in knowledge by exploring the ways in which scientific research datasets that are publicly shared have been reused. Specifically, I examine use requests from three biomedical data repositories in order to answer questions about who is reusing these datasets, how they are using them, and why some datasets are used more than others.

1.1 Background of the Research

A number of cultural and policy changes in the last few years have increased the availability of scientific research data for reuse. In 2013, the United States Office of Science and Technology Policy (OSTP) issued a memo directing agencies to develop policies to increase public access to research data generated using federal funds (Holdren, 2013). Accordingly, federal funders including the National Science Foundation (NSF) and National Institutes of Health (NIH) have created policies requiring researchers to share their data (National Institutes of Health Office of Extramural Research, 2016; National Institutes of Health Office of Science Policy, 2017; National Science Foundation, 2010). The International Committee of Medical Journal Editors (ICMJE) has encouraged member journals to require that authors make data underlying their articles publicly available (Taichman et al., 2017), and many major publishers have already done so, including PLoS and Nature (Nature

Publishing Group, 2017; Silva, 2014). Researchers themselves are also increasingly embracing a culture of greater data sharing and transparency under the umbrella of various open science practices (Nosek et al., 2015b).

As previous research has demonstrated, data sharing and openness bring a number of benefits to the researchers who share, the scientific community, and the general public. In addition to enhancing scientific reproducibility (Ioannidis, 2014; Munafò et al., 2017), shared data can be reused by other researchers, potentially to answer new questions not addressed in the original research. Data reuse increases the return on the investment of the original grant, and also saves on funding that would have been used to gather new data (Arzberger et al., 2004; Costello, 2009). The speed of scientific discovery, and in turn translation to clinical practice, can be accelerated when researchers can reuse existing data instead of spending months or years collecting new data (Knoppers, 2014; Knoppers, Harris, Budin, & Edward, 2014). Researchers who share their data may be rewarded in the form of increased citations to articles with associated publicly available data, as well as opportunities to collaborate and co-author publications with the researchers who reuse their shared data (Piwowar, Day, & Fridsma, 2007; Tenopir et al., 2015).

Despite the potential benefits of reuse and the increasing availability of data, research on the actual impacts of data reuse is so far sparse. Some studies have considered patterns of data request and citation for individual repositories (Coady et al., 2017; Paltoo et al., 2014), but less research has been done to gain a deeper understanding of the impacts that shared research data can have, as well as to

determine how to quantify or measure that impact. Such research has important implications for both policy and practice. Data sharing policies should be founded on a strong evidence base that demonstrates the impacts and benefits of data sharing (Pryor, 2009). The time and effort required to share and curate data is not trivial (Leonelli, 2014), so quantifying the actual impacts of these datasets – as well as determining which datasets have the most potential for long-term impact – helps assure that these investments are worthwhile. Understanding how and why researchers reuse data could also inform development of better technical infrastructure to facilitate discoverability and enhance reuse (Jagodnik et al., 2017). Finally, understanding patterns of data reuse could incentivize sharing by making it possible to build upon existing academic reward structures to give credit to researchers who share high-use and high-impact datasets (Olfson, Wall, & Blanco, 2017). At present, most academic institutions do not recognize shared data as a scholarly product in the context of tenure and promotion decisions, likely because tracking data reuse is technologically challenging and the impact on the broader scientific community of shared datasets is difficult to quantify (Ali-Khan, Harris, & Gold, 2017; Piwowar, Becich, Bilofsky, Crowley, & on behalf of the caBIG Data Sharing and Intellectual Capital Workspace, 2008).

While tracking data reuse across science in general may be informative, the question of how to quantify data reuse and its impacts is especially salient in the context of biomedical research. In some disciplines, such as geology and astronomy, a culture of data reuse is relatively well established, given that these research

communities have a long history of sharing data generated by a small number of sensors or telescopes, which is then analyzed by researchers around the world (Giles, 1995; Pepe, Goodman, Muench, Crosas, & Erdmann, 2014). However, widespread data sharing and reuse has not been the norm in biomedical research, and biomedical researchers have expressed both less willingness to share their own data and less interest in using others' data (Tenopir et al., 2011, 2015). Some biomedical researchers even consider data reuse anathema, with one controversial editorial decrying researchers who reuse data as "research parasites" (Longo & Drazen, 2016).

One argument of its detractors is that sharing data will discourage researchers from undertaking large studies, particularly clinical trials, because they expect to be able to publish multiple articles over the course of several years using the data (The International Consortium of Investigators for Fairness in Trial Data Sharing, 2016). Sharing the data before they have the chance to conduct longer term studies, they argue, means that other researchers could "scoop" them – beat them to publication on discoveries that they could have gotten credit for. Given that articles are one of the most important currencies in academic credit systems, this argument suggests that identifying a means to reward researchers for sharing data could alleviate some of these concerns and remove some of the disincentives to sharing. Indeed, the NIH's recent Strategic Plan for Data Science recognizes that "appropriate reward…systems are central to making data FAIR [findable, accessible, interoperable, and reusable] and for incentivizing researchers to share their data and analysis tools widely for reuse by others" (National Institutes of Health, 2018b, p. 24).

While the research presented here does not necessarily solve the deeper cultural problems associated with biomedical data sharing, the findings of this study will help lay the foundation for solutions by providing a deeper understanding of the nature of biomedical data reuse. Data sharing cannot be meaningfully rewarded, nor can informed decisions be made about data curation and preservation, if it remains unclear how much datasets are being reused, who is reusing them, and for what purpose. This study explores biomedical data reuse in ways that will help answer these questions, as well as providing insight into how repositories and funders can make evidence-based decisions about policy and practice.

1.2 Research Questions

To better understand how and why biomedical researchers reuse existing datasets, this dissertation is guided by four research questions:

Research Question 1: What are the purposes and characteristics of biomedical research reuse?

> **Research Question 1.1:** For what methods and analysis types are datasets reused?
>
> > **Hypothesis 1.1:** Genomic datasets of the type found in dbGaP will be more likely to be used in combination in meta-analyses, while clinical datasets of the type found in the NIDDK repository will be more likely to be used on their own to answer an original research question.

Research Question 1.2: How closely are the topics for data reuse aligned with the topics for which the data were originally collected?

Hypothesis 1.2: Similarity between original topics and topics of reuse will be lower for genomic data (found in dbGaP) than for clinical data (found in the NIDDK repository).

Research Question 2: What are the demographics of researchers who reuse existing datasets?

Research Question 2.1: Where are requestors located in the world?

Hypothesis 2.1: Requestors will be primarily located in regions with a greater proportion of research institutions, including North America, Europe, and Asia.

Research Question 2.2: Are there patterns in career stage of requestors?

Hypothesis 2.2: A broad range of career stages, from student to full professor (or equivalent) will be represented.

Research Question 3: Are there temporal patterns to dataset requests?

Hypothesis 3: Patterns of requests relative to the original dataset release date will demonstrate a cumulative advantage process, similar to other scientific communication processes such as article citation.

Research Question 4: Are there dataset topics that are more highly requested?

These four questions approach the topic of reuse from two perspectives. Research Questions 1 and 2 answer questions about the characteristics of requests and requestors: who are the requestors and what are they planning to do with the data?

Research Questions 3 and 4, on the other hand, examine characteristics of the datasets: which datasets are most requested and how does a dataset's requests evolve over the years after its release? Together, the findings of these questions will provide a better understanding the complex phenomenon of biomedical data reuse.

1.3 Scope of this Study

Broadly speaking, "biomedical research data" can include many different types of data generated, collected, or used in the course of the wide range of research activities that biomedical researchers conduct. In its original 2003 statement on data sharing, the NIH specifically notes that only "final research data" fall within the purview of its sharing policy. Their definition of final research data is the "recorded factual material commonly accepted in the scientific community as necessary to validate research findings." They further note that other research objects such as "laboratory notebooks, partial datasets, preliminary analyses, drafts of scientific papers, plans for future research, peer review reports, communications with colleagues, or physical objects, such as gels or laboratory specimens" do not constitute final research data and are therefore excluded from policies regarding sharing (National Institutes of Health Office of Extramural Research, 2004).

The 2003 statement also recognizes that there are many mechanisms by which research data may be shared, from the relatively restrictive (interested parties must contact the original researcher to negotiate access) to the maximally open (data are made freely available in a public repository). In more recent policies and mandates

from publishers and funders, the once-acceptable "data available upon request" is often considered inadequate as a means of sharing, especially since requestors have often found that authors cannot or do not share data upon request (Langille, Ravel, & Fricke, 2018; Savage, Vickers, Kats, & Molenaar, 2009; Stodden, Seiler, & Ma, 2018). Instead, most policies encourage, and sometimes even require, researchers to make data freely available in a repository, although that ideal is not always fully realized (Federer et al., 2018). Given the policy move toward repositories as the "gold standard" for data sharing, this study focuses on data shared within public biomedical data repositories. This choice is also based on practical considerations; data kept within an individual researcher's lab would not only be difficult for someone else to reuse, but nearly impossible to identify for inclusion in this study.

A further challenge to this research is identifying means for quantifying data reuse. Obtaining accurate counts of reuse of research datasets is challenging, given that standards for data citation have not been widely adopted yet. My previous research on the correlation between data use requests and citations to those datasets in the published literature found that the average dataset from biomedical data repositories had between about five and nine use requests for every one citation, suggesting that most use requests do not result in a publication that can be identified using existing search tools (Federer, 2018). While many open repositories track download counts for datasets, such raw counts provide little insight into who is using the data and for what purpose, or even whether they end up actually using the data at all.

In the absence of a tool or method for accurately quantifying and tracking reuse of shared datasets, this study utilizes use requests submitted for controlled-access biomedical datasets as a proxy for data reuse. This study considers three repositories administered by various groups within the NIH, all of which make their use requests publicly available. The Database of Genotypes and Phenotypes (dbGaP), housed at the National Center for Biotechnology Information (NCBI), contains human genetic sequence data and associated diseases or characteristics (National Center for Biotechnology Information, 2018). The BioLINCC repository and the NIDDK Central Repository contain datasets arising from research funded by the National Heart, Lung, and Blood Institute (NHLBI) and the National Institute of Diabetes and Digestive and Kidney Diseases' (NIDDK), respectively (National Heart, Lung, and Blood Institute, 2018; National Institute of Diabetes and Digestive and Kidney Diseases, 2018). Together, these three repositories cover a range of data types, from clinical data (NIDDK and NHLBI) to genomic data (dbGaP), as well as a range of diseases and topics.

As will be further discussed throughout this dissertation, this method for operationalizing reuse has certain limitations, as does the selection of these particular repositories. A request for a dataset does not necessarily guarantee that the requestor ended up using it, nor can it be known for certain whether the person who requested the data was the person who intended to use it – for example, a professor might request a dataset on behalf of a student. Still, these use requests provide a richer source of information about how biomedical datasets are reused than other currently

available methods. Throughout this study, I will note how the methodologies and data sources used here limit the generalizability of these results and provide specific discussion about how these results can be meaningfully and responsibly applied.

1.4 Study Methodologies

This study utilizes a mixed methods approach, combining qualitative and quantitative methods to gain a holistic view of the reuse of biomedical datasets from the three repositories. Some of this work considers the requests and requestors, while other parts of the study focus on the datasets themselves. Taken together, these different pieces of data and types of analysis form a view of the who, what, when, where, and why of data reuse.

In the first part of the study, content analysis of the use requests provides insight into *who* is making requests, *where* in the world they are located, and *why* they would like to reuse the data. I coded use requests for the type of reuse using a taxonomy drawn from the literature and inductively expanded to address types of reuse not previously identified. This analysis provides insight into the ways different types of data are reused. Using an automated indexing tool, I further coded requests with topics drawn from a controlled vocabulary that the repositories also use to describe the datasets. Comparing the similarity between topics in the requests to topics in the datasets provides a quantitative means to understand how similar intended data reuse is to the reasons for which the data were originally collected. By analyzing demographic information about the researchers who request datasets, this

study also provides an understanding of who is benefitting from shared data – specifically, what is the career status of researchers who request data, and where in the world are they geographically located?

The second part of the study focuses on analysis of the patterns of reuse of the datasets, investigating *when* in the data's life cycle it is requested and *what* topics are most requested. Analyzing patterns of requests over the course of a dataset's life can yield insight into the long-term usefulness of a dataset, as well as provide an understanding of how similar patterns of request are to other processes in science, such as citations to articles over time. I also conducted text mining to determine whether there are topics that are more highly requested than others. Using topic modeling on repository-provided dataset descriptions yields groupings of datasets that are conceptually similar. Examining patterns of reuse among those topics enables identification of highly requested topics. Understanding these "when" and "what" questions of data reuse could aid in early identification of datasets that will go on to be highly requested; datasets that show early signs of high reuse patterns or those that cover highly requested topics could be prioritized for more in-depth curation.

1.5 Importance and Contributions

The findings of this study will have implications for a number of different stakeholders interested in how to track and quantify data reuse. At present, rewarding researchers for sharing data is challenging because of the difficulties in identifying and tracking reuse; moreover, the practical impact of a shared dataset cannot be

quantified. Methods for evaluating research impact generally rely on well-established metrics with widely agreed-upon significance across scholarly communities. For example, the impact of an article may be quantified by the number of times other researchers cite it; the impact of a research grant may be quantified by the number of patents it generates or the market value of a drug that it yields. While these measures may be imperfect representations of the practical impact of a researcher's work and productivity, they still represent a common currency used in the context of tenure, promotion, retention, and funding decisions (Carpenter, Cone, & Sarli, 2014; Holden, Rosenberg, & Barker, 1994; Moher et al., 2018).

If data sharing is to be rewarded, the research community must come to consensus about how the impact of a shared dataset is quantified. Simple counts of use requests or downloads elide the many, often very different, forms of reuse. By providing a better understanding of how datasets are reused, this research will help inform how to most effectively and fairly reward data sharing. Thus, these findings may provide insight for funders that wish to reward researchers for sharing, for academic institutions that want ways to measure the impact of their researchers' contributions, and for researchers who often spend significant time and effort to share data but do not yet have mechanisms to be rewarded for doing so (CODATA-ICSTI Task Group on Data Citation Standards and Practices, 2013; Mooney & Newton, 2012).

This understanding of how to quantify the impact of data sharing has further implications for development of policies informing data sharing and reuse. The

aforementioned OSTP memo calling for federal agencies to create policies to increase public access to research results, including data, was issued in 2013; five years on, policy in these areas is still nascent (Holdren, 2013). For example, as of this writing, the NIH has yet to issue such a policy, although they have received and are reviewing public comments on a Proposed Provisions for a Draft Data Management and Sharing Policy (National Institutes of Health, 2018c). The NIH's existing policy only requires a data sharing plan for grants of over $500,000 annually and does not consider the content of the plan in its competitive review process (National Institutes of Health Office of Extramural Research, 2004). A better understanding of the ways that shared data contribute to advancing science through reuse could help inform future policy developments.

Academic research institutions, too, are beginning to adopt policies that could be informed by the results of this study. For example, the Montreal Neuroscience Institute (MNI) has adopted an institution-wide open science policy that includes rewarding open sharing of data and other research products in the tenure and promotion process (Ali-Khan, Jean, MacDonald, & Gold, 2018). At the same time, they have recognized that doing so requires understanding how to quantify the impact of open science products and are developing a toolkit of qualitative and quantitative techniques to do so (Gold et al., 2018). That work, which has drawn on the input of international experts in open science, will be further enhanced by the deeper understanding of data reuse that this study will yield.

This study also has implications for the repositories that host data and the curators who do the often time-consuming work of making data ready for reuse (Leonelli, 2014; Levin & Leonelli, 2017). Traditional libraries must make choices about which materials they will commit to preserve and which they will discard because physical space and other resources are limited. Similarly, it is neither feasible nor desirable to curate and preserve every single research dataset in perpetuity. It is difficult to predict which datasets will have future value, in part because biomedical research is a moving target – the hottest topics and most advanced technologies of today can quickly become outdated – but understanding what characteristics are most predictive of reuse can build an evidence base for making well-informed decisions about which datasets to prioritize.

By exploring the demographics of the researchers who reuse datasets, this research may also provide a better understanding of how data sharing can help democratize science and facilitate research in areas where funding resources are sparser. For example, in regions of the world where less scientific funding is available, generating certain types or large quantities of data may be financially out of reach (Serwadda, Ndebele, Grabowski, Bajunirwe, & Wanyenze, 2018). Even in countries where research is comparatively well funded, resources may not be equally distributed. Early career researchers, women, under-represented minorities, and researchers at smaller institutions may not have the resources or funding that are required to generate certain types of data. Not every researcher has access to sophisticated high-throughput sequencing machines or a cadre of staff to collect

years' worth of longitudinal data. If these data already exist, sharing them may be a more efficient way to distribute limited resources while maximizing scientific discovery. By better understanding who is (and who is not) currently using shared data resources, this research will be useful to funders who may wish to fund research that encourages reuse of existing resources, as well as to repositories and others who may be in a position to conduct outreach to increase awareness of the availability of such resources.

1.6 Organization of the Dissertation

This dissertation comprises seven chapters, including this introduction. Chapter 2 reviews the literature to contextualize this research within the literatures of science and technology studies, open science, and scholarly metrics. Chapter 3 describes the design of this study, including discussion of methods for data collection and analysis. The findings of these analyses are split into two chapters; Chapter 4 describes findings based on analysis of use requests and requestors, while Chapter 5 focuses on the analysis of the datasets themselves. Chapter 6 synthesizes these findings to better define the who, what, when, where, and why of biomedical data reuse. Finally, Chapter 7 discusses the implications of these findings for various stakeholders in the biomedical research community and outlines directions for future research that builds on this exploratory study.

Chapter 2: Review of the Literature

Although the wide availability of research data for reuse is a relatively new development, various areas of inquiries into scholarly communication and academic reward systems, as well as researchers' data use and reuse behaviors, provide a foundation for this study. This chapter begins with a discussion of research impact, including the historical context of how and why research impacts are measured, as well as an examination of how these measures are used in the context of academic reward systems today. Understanding how research outputs are currently measured and rewarded backgrounds this study's approach to how metrics of data reuse could fit within existing scientific reward structures, and therefore provides insight into what characteristics of datasets and data reuse should be considered in a model for quantifying the impact of datasets. Some of these approaches draw on established bibliometrics techniques; although citations to articles cannot be considered exactly equivalent to instances of data reuse, many of the approaches used in the context of articles can yield insight into the quantification of data reuse.

This chapter also draws upon the nascent literature on data sharing and reuse to provide background on what is already known about how researchers reuse data. Many of these studies consider scientific research from non-biomedical fields; while it has been established that different disciplines have different cultures of data sharing and reuse (Tenopir et al., 2011, 2015), these studies provide important ideas about how to conceptualize data reuse and its role in advancing science.

2.1 Scientific Credit and Reward

Before attempting to track, quantify, and predict data reuse, it is essential to understand the ecosystem of credit and reward within which science operates. At the heart of science, of course, is the attempt to understand the phenomena that drive the world around us, but the pursuit of knowledge is arguably not the only goal of many researchers – rather, it is the pursuit of knowledge that will allow them to gain credit in the scientific community.

2.1.1 The Role of Credit in Science

Robert K. Merton has posited "four sets of institutional imperatives taken to comprise the ethos of modern science": communalism, universalism, disinterestedness, and organized skepticism (Merton, 1942, p. 270). The norm of disinterestedness suggests that science be conducted for the common good rather than the researcher's personal benefit, particularly in the context of financial gain. By communalism, Merton means that scientific knowledge should be "owned" communally by, and therefore be accessible to, the entire scientific community in order to facilitate collaboration and advance research. This argument is especially salient in the context of federally funded research; as the Office of Science and Technology Policy's 2013 memorandum on Increasing Access to the Results of Federally Funded Scientific Research points out, the outcomes of research should be available to the public that has funded it through their tax dollars (Holdren, 2013).

Some critics have argued that Merton's norms do not present a comprehensive view of the normative structure of science, suggesting that "counternorms" often drive scientists' behavior and serve a function in scientific communities. For example, scientists regularly engage in secrecy, the counternorm to communalism, by strategically withholding information to ensure that others cannot steal credit for their work. Some secrecy is probably essential to the social structure of science, as without it, "science would degenerate into a state of continual warfare" (Mitroff, 1974, p. 593).

Like Anderson et al., I suggest that Merton's norms are best viewed as "ideals that...are counterbalanced by opposing norms" (Anderson, Ronning, DeVries, & Martinson, 2010, p. 5). Scientific knowledge progresses most effectively when researchers operate somewhere between complete secrecy and complete openness, in a system that provides them with a mechanism for receiving credit for their contributions while still allowing them to build upon the knowledge of others. This view is not incompatible with Merton's norms – despite arguing for a high level of openness and community ownership of knowledge, Merton does not suggest scientists should work without reward or acknowledgement. Rather, he suggests that "the scientist's claim to 'his' intellectual 'property' is limited to that of recognition and esteem," and argues that, when scientific institutions function well, they reward scientists proportionally to the significance of their work (Merton, 1942, p. 273).

Article citation is an essential mechanism for enabling this proportional reward process. The practice of citing articles makes it possible to trace influence and

inspiration and serves the very practical purpose of giving credit to researchers for their scholarly labor. Researchers need not pay a licensing fee or purchase an idea to build upon it in their own work; the "payment" for the idea is rendered to the original creator in the form of a citation. A citation on its own has no monetary value, but citations have very real economic impacts on researchers, given that they are often used in academic hiring, tenure and promotion, and funding decisions (Carpenter et al., 2014; Durieux & Gevenois, 2010; Holden et al., 1994).

Several researchers have explored the concept of credit and its important role in the economy of the research community. In their seminal work *Laboratory Life: The Construction of Scientific Facts*, Latour and Woolgar devote an entire chapter to "Cycles of Credit" (1986). They describe science as a process of accumulating credibility capital through recognition in the form of citations, awards, and credentials, which can in turn be "reinvested" to receive the necessary resources to continue conducting research, such as grant funding, laboratory resources, and tenure. "The notion of credibility," they argue, "makes possible the conversion between money, data, prestige, credentials, problem areas, argument, papers, and so on" (Latour & Woolgar, 1986, p. 200). Other scholars have also taken an economic view of the function of credit and citation, for example, describing citation as payment of an "intellectual debt" (Garfield, 2002; Kochen, 1987). Merton argues that, in a sense, getting citations is the impetus for scientific publishing in the first place, pointing out that, "since recognition by qualified peers is the basic form of extrinsic reward…and since that reward can be accorded only when the work is made known, this

historically evolving reward system provides institutionalized incentive for open publication without direct financial reward" (Merton, 1983, para. 5).

The importance of credit in scientific research is underscored by the grave tone of discussions about instances in which credit is not properly given. Variously termed as "citation amnesia," "bibliographic negligence," "disregard syndrome," and "petty larceny plagiarism," the failure of a researcher to cite an article that has informed his or her work is considered a serious breach of scientific conduct (Garfield, 1982, 1991; Ginsburg, 2001; Maes, 2015). Lack of proper attribution is also framed as a moral failing, with citation being described as "a matter of science's family values" and failing to cite as "a menace to honest science" and a "serious transgression" (Garfield, 1991; Ginsburg, 2001; Palevitz, 1997). Garfield even muses on the possibility of establishing a "science court" that would enforce the norms of citation and "met[e] out punishment to willful perpetrators" (Garfield, 1987, 1989, 1991, para. 2).

In light of this economy of reward and the intellectual theft that failure to cite represents to the scientific community, it is not difficult to understand why some researchers would be unwilling to share their data. As I will discuss, standard mechanisms for researchers to cite datasets they have reused have not yet been widely adopted. Data sharing detractors see data reuse as "possibly stealing from the research productivity planned by the data gatherers" (Longo & Drazen, 2016, para. 3). From a game theory perspective, though sharing is good for the community at large, researchers would not logically do so, since "there is a conflicting interest for

individual researchers, who are always better off not sharing and omitting the sharing cost while they would have higher impact when sharing as a community" (Pronk, Wiersma, van Weerden, & Schieving, 2015, p. 1). Until mechanisms exist to situate data reuse within the scientific economy – that is, to quantify data reuse and reward researchers who share high-value datasets that go on to be frequently reused – many researchers may see reuse of their data as intellectual theft.

2.1.2 Metrics for Scientific Credit

The assumption that citations equal credit is foundational to the field of bibliometrics, which has at its aim measuring scientific impact. Bibliometricians use various indicators and statistical methods to assess the value of articles, the impact of journals, and the productivity of researchers. For example, the h-index, calculated by considering the number of citations for each paper in a researcher's body of work, is often used in hiring and funding decisions and has been demonstrated to be effective in comparing researchers' outputs and predicting future scientific success (Acuna, Allesina, & Kording, 2012; Bornmann & Daniel, 2007; Carpenter et al., 2014; Hirsch, 2005, 2007; Penner, Pan, Petersen, Kaski, & Fortunato, 2013).

Despite the widespread use of bibliometric methods, uncertainty remains about how well these measures accurately reflect scientific achievement and productivity. Article citations are not always easy to collect, and analysis may provide incomplete results (Lane, 2010). Article citation is also only one means of measuring scientific output, and cannot capture uses of scientific knowledge that occur outside

of the traditional scientific literature (Hicks, Wouters, Waltman, de Rijcke, & Rafols, 2015). As Priem puts it, "ideas do not leave good tracks" (Priem, 2014, p. 263). As access to articles has become largely digital, new methods for counting article use have emerged, including article downloads, Mendeley readership, and mentions in online sources such as blogs and social media, but their validity and significance remains unclear (Bollen, Van De Sompel, Smith, & Luce, 2005; Galligan & Dyas-Correia, 2013; Schlögl, Gorraiz, Gumpenberger, Jack, & Kraker, 2014; Thelwall, Haustein, Larivière, & Sugimoto, 2013).

To address some of these limitations, bibliometrics researchers have undertaken research to consider how effectively article citation reflects impact. Using citations as a means to reward impactful science assumes that citations are positive, though in Garfield's influential list of fifteen reasons for citations, some are actually negative, such as "criticizing previous work" or "disclaiming work or ideas of others" (Garfield, 1964, p. 85). A growing body of literature explores what citation counts *actually* measure, including quantitative studies of citations and qualitative studies of researchers' citing behaviors (Bornmann & Daniel, 2008). Although significant questions remain about researchers' motivations for citing articles, citation counts are still widely recognized as a "a strong indicator of scientific performance" (van Raan, 2005, p. 3).

Nonetheless, even when bibliometric measures can be relatively well defined and easily measured, bibliometricians urge caution in interpreting and using these metrics for decision-making. The 2015 "Leiden Manifesto," a declaration of best

practices for bibliometrics, described a scientific community inundated by metrics that are "usually well intentioned, not always well informed, often ill applied" (Hicks et al., 2015, p. 429). They warn that citation counts in particular are subject to "conceptual ambiguity and random variability" and urge the scientific community to "avoid misplaced concreteness and false precision" when interpreting and using all types of research impact measures (Hicks et al., 2015, p. 431).

In considering how to quantify data reuse and measure its impact, an important caution that researchers should take from bibliometricians is to consider potential unintended consequences. Some critics see citation counts as an example of Goodhart's Law, which states that "when a measure becomes a target, it ceases to be a good measure" (Edwards & Roy, 2017, p. 52). They argue that using article citation to reward high-impact science may have the effect of creating perverse incentives that encourage self-citation and other bad behaviors that artificially inflate citation counts (Edwards & Roy, 2017; Werner, 2015). Further, it is also essential to consider how to measure what is *meaningful*, and not simply what is easy to count, especially in the context of phenomena for which there is an "absence of internationally meaningful comparative data" – a situation that is almost certainly the case for data reuse at present (Hazelkorn, 2013, p. 6). As the "Leiden Manifesto" points out, "the problem is that evaluation is now led by the data rather than by judgement" (Hicks et al., 2015, p. 429).

2.1.3 Patterns of Scientific Attention

Merton has described a social phenomenon in science that he termed "the Matthew effect" after a parable that states "for to every one who has will more be given, and he will have abundance; but from him who has not, even what he has will be taken away" (1968, p. 159). He describes this effect at the level of the investigator, suggesting that "the accruing of greater increments of recognition for particular scientific contributions to scientists of considerable repute and the withholding of such recognition from scientists who have not yet made their mark" (Merton, 1968, p. 159). In other words, the more well known a researcher is, the more likely he or she is to gain further attention. Bibliometric research has demonstrated that this effect, also called the "success breeds success" phenomenon (Cozzens, 1985), exists in article citations as well; that is, articles that are highly cited are more likely to receive more citations in the future (Bornmann & Daniel, 2008; Burrell, 2003; Cozzens, 1985).

Given that this phenomenon occurs at the researcher level as well as the article level, it stands to reason that dataset reuse may also governed by such a model, such that the more a dataset is reused, the more attention it gets and the more likely it is to be reused. Further, it has been shown that the data creator's reputation is a factor in a researcher's decision to use a dataset (Faniel, Kriesberg, & Yakel, 2015). Since researcher reputation is subject to the Matthew effect, it follows that the success of the researcher will breed success of his or her datasets. In the context of data science, this process is likely especially true for benchmarking datasets, which are used for testing new tools and methods, as well as comparing them to existing gold standard

tools (Moura et al., 2013; Ó Conchúir et al., 2015). Datasets that have been used for benchmarking are more likely to go on to be used for this purpose again, since it is useful to compare a new tool to an existing tool on the same dataset.

In statistics, the Matthew effect is described as a cumulative advantage process (de Solla Price, 1976). Some bibliometricians argue that citations to papers are accrued in a linear fashion at a constant rate (Bornmann & Daniel, 2008; Hirsch, 2005). Others contend that papers accrue citations at random, therefore arguing for a stochastic model (Burrell, 2003, 2008). De Solla Price suggests that accumulated citations are determined by the number of citations that the article receives early, which he terms the initial pulse (1976). These types of models of cumulative advantage may be helpful for predicting future reuse of datasets.

While dataset reuse likely follows some of the same patterns of article citation over shorter time spans, the long-term patterns may differ. Even the most highly cited papers are subject to a process of "attention decay;" citations hit a peak, typically between two and seven years depending on discipline, and citations subsequently taper off (Eom & Fortunato, 2011; Parolo et al., 2015). Attention decay in articles is largely driven by knowledge obsolescence; as new discoveries are made and new articles written, researchers are more likely to cite the newer, more current information (Fortunato et al., 2018). However, this same process may not hold with datasets. Some of the datasets considered in this study demonstrate that even old data can be of significance to researchers; for example, in the NHLBI repository, over 20% of the datasets were collected more than 20 years ago. That these datasets are

still being requested suggests that datasets may not be subject to the same pattern of attention decay as articles.

2.2 Scientific Data Sharing and Reuse

The sharing of data with other researchers is not new to science. However, data previously tended to be shared through interpersonal connections, such that a researcher who wanted to use a dataset had to first know that it existed, and then negotiate with the data creator for access. This process requires significant and often tacit knowledge of the discipline and interpersonal connections within the field, limiting opportunities to students, early career-researchers, under-represented minorities, and others who were not research community insiders (Wallis, Rolando, & Borgman, 2013; Yoon, 2017; Zimmerman, 2007). Data was exchanged in the context of a "gift economy" – sharing in an open repository would be undesirable because data had value as an item to be bartered with other researchers in return for resources or intangible credit capital (Bollen, Van de Sompel, Hagberg, & Chute, 2009; Wallis et al., 2013). Data use agreements governed how reusers would be expected to "compensate" the data sharer, sometimes in the form of co-authorship on resulting papers (Gorgolewski, Margulies, & Milham, 2013). Although this type of sharing still occurs, the development of computational and technological infrastructure that has enabled the creation of data repositories, as well as the policy mandates that have driven researchers to populate them, have inherently changed how datasets are shared today (Tausczik, 2016).

2.2.1 Understanding Data Reuse

Given that widespread data availability is a relatively new phenomenon, our understanding of how researchers are reusing publicly available datasets is still emerging. Coady et al. (2017) developed a set of categories for coding reuse requests in their study of the NHLBI repository (emphasis added for ease of reading):

> **new question**, defined as a secondary analysis designed to explore associations, prognostic factors, subgroup analyses, or similar issues; **meta-analysis or pooled study**, defined as a formal meta-analysis of individual participant data, combined study analysis, or consortium of studies with participant-level data; **statistical methods**, defined as a project focused on the development and testing of new statistical approaches; **clinical trial methods**, defined as a project examining statistical methods or analytic approaches that are generalizable to all or specific types of clinical trials; and **other projects**, examples of which include pilot data for a subsequent grant submission, simulation studies, and development of prediction equations. (p. 1851)

This taxonomy provides a useful starting point for understanding types of data reuse, although the datasets considered in the Coady et al. study are limited to clinical trial data; additional types of reuse not covered in this taxonomy have been discussed in other studies. For example, publicly available data can be useful in reproducing or verifying the results of an original study, a particularly compelling use given that many scientific disciplines are troubled by a "reproducibility crisis" (Borgman, 2011; Pasquetto, Randles, & Borgman, 2017). As data science methodologies advance,

researchers also need existing data for development and validation of new software, particularly in the context of supervised machine learning tasks, which require well-described and tagged data from which the algorithm can learn patterns (Kotsiantis, 2007). Even beyond research, shared data can have important applications in training the next generation of researchers who do not yet have their own data to analyze. For example, Compute Canada funds cloud-based data and compute hubs for training use in Canadian academic institutions (Compute Canada, 2018).

Some of the variation in ways datasets are reused is due to characteristics of the data themselves. Biomedical data can include a wide range of data types; the two considered in this study, clinical and genomic data, have very different histories that influence the ways they are collected, and therefore the ways they can be reused. Clinical research traces its history back hundreds of years (Bhatt, 2010); by comparison, genomic research is quite young, beginning with the Human Genome Project (HGP) in the 1990s (National Human Genome Research Institute, 2012). Data sharing has been a norm in genomic research from the start – the HGP considered "rapid prepublication data release" fundamental to genomic research, and this principle was even codified in the form of the Bermuda Principles and adopted into policy by the National Institutes of Health (Collins, Morgan, & Patrinos, 2003, p. 288; Powledge, 2003). That type of widespread sharing and collaboration has not been part of the culture of clinical research, which likely contributes to the resistance among many clinical researchers to policies that would require them to share (The International Consortium of Investigators for Fairness in Trial Data Sharing, 2016).

Since genomic research has embraced sharing and collaboration from its beginnings, genomic data have intentionally been standardized; the Genomic Standards Consortium was formed in 2005 to develop and promote data standards (Field et al., 2011). These standards enable researchers not only to use data from another lab, but to aggregate it with their own, which is especially important given that genomic research requires a much larger sample of participants to achieve statistical power than does clinical research (Hong & Park, 2012). On the other hand, little standardization exists across clinical datasets; researchers often word questions to patients in different ways or record the same concept using different terminology (Richesson & Nadkarni, 2011). As a result, even if clinical researchers share their data, other researchers' ability to aggregate it with other datasets is limited. Efforts are underway to improve standardization of clinical data; for example, the National Institutes of Health's activities to promote Common Data Elements would help ensure greater consistency across clinical datasets and thereby enable aggregation and potentially increase reuse (Sheehan et al., 2016).

Beyond the *what* of data reuse, a number of studies have considered the *why*, exploring researchers' attitudes toward and experiences with reusing research data. Tenopir et al.'s 2011 article and their 2015 follow-up provide useful insight into how practices have changed over time. Eighty-three percent of respondents strongly or somewhat agreed that they "would use other researchers' datasets if their datasets were easily accessible" (Tenopir et al., 2011, p. 8). They do not report the percentages for responses in the follow-up article, but do indicate that the agreement with this

statement increased significantly from a mean of 4.19 to 4.33, on scale of 1 (disagree strongly) to 5 (agree strongly) (Tenopir et al., 2015). However, these attitudes differ across disciplines; notably, researchers in medical and health sciences fields had the lowest rate of agreement with the statement in both the original study and its follow up (Tenopir et al., 2011, 2015).

Other studies have aimed to understand the reasons underlying researchers' attitudes about reuse. Several studies have found that trust plays a major role in researchers' decision to reuse data (Faniel & Jacobsen, 2010; Faniel et al., 2015; Rolland & Lee, 2013; Yakel, Faniel, Kriesberg, & Yoon, 2013; Yoon, 2014, 2017), although another study found that reuse decisions were more based on perceived usefulness of the data than its trustworthiness (Kim & Yoon, 2017) . The concept of trustworthiness may be tied to the repository (does the researcher trust the repository to curate, preserve, and provide accurate data?), as well as the original data collector (does the data collector have a reputation for accurate and clean data?). Characteristics of the datasets themselves also play a significant role in researchers' selection of datasets to reuse. Researchers look for datasets that are complete, credible, accompanied by high-quality metadata, and easy to use (Faniel & Jacobsen, 2010; Faniel et al., 2015). However, most of these studies considered reuse in specific research disciplines, such as earthquake engineering or social sciences, and little research has interrogated the practices and attitudes of biomedical researchers. Given Tenopir et al.'s (2011, 2015) findings that researchers in biomedical research differ from their counterparts in other disciplines in many ways regarding sharing and reuse,

these findings may not be generalizable to biomedical researchers. Differences likely exist even within specific sub-disciplines of biomedical research; a previous study my colleagues and I conducted found that NIH clinical researchers were significantly less likely to consider data reuse important to their work than non-clinical researchers at NIH (Federer, Lu, Joubert, Welsh, & Brandys, 2015).

Not only are biomedical researchers different from those in other disciplines, but the data used in biomedical research is also different in an important way: it often contains personally identifiable information on human subjects. Data reuse in the context of biomedical research therefore raises some additional concerns about privacy that may not apply to other types of research data. The Health Insurance Portability and Accountability Act of 1996 stipulates that patients' data cannot be shared without their consent, thus limiting the sharing of some types of patient data, unless they can be de-identified adequately to present "only a very small risk" of the patient being re-identified (Meystre et al., 2017). However, in some cases, such as patients with very rare diseases, de-identification may not be possible (Hansson et al., 2016; Wan et al., 2017). Paradoxically, it is these very patients who could potentially stand to benefit the most from data sharing, since collecting enough data to draw statistically meaningful conclusions often necessitates researchers from around the world sharing data on their patients.

2.2.2 Challenges in Tracking and Quantifying Data Reuse

As data reuse becomes more common, many in the scientific community have recognized the need for mechanisms to track and quantify data reuse. One approach that has been championed by various stakeholders is data citation (Bierer, Crosas, & Pierce, 2017). In 2014, the scholarly communication organization FORCE11 issued a Joint Declaration of Data Citation Principles that suggests "data should be considered legitimate, citable products of research," and proposes eight principles for the "purpose, function and attributes of citations" (Data Citation Synthesis Group, 2014). This formal declaration is situated within a body of literature exploring both the ideal forms that data citation might take (Altman & Crosas, 2013; Altman & King, 2007; Silvello, 2017) and actual citation practices observed in the literatures of various disciplines (Edmunds, Pollard, Hole, & Basford, 2012; Henderson & Kotz, 2015; Mooney & Newton, 2012). Advocates see data citation as a means to enhance scientific reproducibility by allowing readers to easily locate data underlying scientific articles (Altman & Crosas, 2013; CODATA-ICSTI Task Group on Data Citation Standards and Practices, 2013). Researchers themselves also seem to consider data citation important: in 2011, 92% of respondents said that they "agree strongly" or "agree somewhat" with the statement "it is important that my data are cited when used by other researchers," although a 2015 follow-up found significantly less agreement with that statement (Tenopir et al., 2011, 2015).

However, utilizing data citations as a means for quantifying reuse remains challenging, especially since standards have not been widely adopted (Zhao, Yan, &

Li, 2017). While article citations are standardized and typically found in a reference section, authors place data citations throughout articles, including in the acknowledgements, materials/methods section, or elsewhere (Callahan, Winnenburg, & Shah, 2018; Piwowar, Carlson, & Vision, 2011). Others may cite not the dataset itself, but an article describing the dataset. For example, the GenBank database directs researchers who have reused data to cite a paper describing the database (Benson, Karsch-Mizrachi, Lipman, Ostell, & Wheeler, 2005). Citations to that paper do not necessarily reflect use of GenBank data; authors may cite that paper even when GenBank data have not been used (such as I have done here).

Inconsistencies in data citations complicate the process of locating articles that report on reuse of a dataset. A variety of academic databases have article citation indices that automatically connect a user to citing articles, but a similarly comprehensive data citation does not yet exist (Garfield, 1955; Robinson-García, Jiménez-Contreras, & Torres-Salinas, 2015). Though some computational and automatic methods have been developed (Piwowar, 2010; Q. Zhang, Cheng, Huang, & Lu, 2016), correctly and completely identifying articles citing datasets often requires significant manual work. Most studies that have utilized citation-based methods to quantify data reuse have relied at least partly on manual identification of articles and elimination of false positives (Belter, 2014; Callahan et al., 2018; Piwowar et al., 2011). These methods would be impractical in large-scale analyses to systematically quantify the impact of larger sets of data citations. In my previous research, I demonstrated that some articles that report on data reuse do not cite the

original dataset at all; I could include them in my study because the repository containing the cited dataset had been notified of the publication and so included it in the list they provided me. Had I attempted to locate all articles citing those datasets on my own, it would have been impossible for me to identify them (Federer, 2018).

Despite the slow uptake, it appears that many within the scholarly community are eager to move toward data citation standards and infrastructure that allow for better tracking of data and its reuse in the scholarly literature. These efforts could also be stimulated by increased recognition of data reuse as a form of scientific impact that merits scholarly credit.

2.3 Conclusions

Although the study of data reuse is relatively young and many questions remain about who is reusing biomedical research data and for what purposes, the bodies of research described here can help inform directions for this research. Quantifying and tracking data reuse is important for ensuring proper credit and attribution. Because data reuse is situated within the context of an existing structure for academic credit, this research is most useful if it builds upon our current understanding of how researchers interact with and use scientific knowledge of all types. As the next chapter will discuss, applying bibliometric models of understanding scientific credit and reward also supplies a useful set of methods for quantifying and tracking data reuse.

Chapter 3: Methodology

Chapters 1 and 2 have provided an overview of the need for and challenges involved in quantifying and understanding biomedical data reuse. In this chapter, I will describe the research design and the approach I have taken to answer four research questions intended to elucidate who is reusing data and how they are doing so:

Research Question 1: What are the purposes and characteristics of biomedical research data reuse?

Research Question 1.1: For what methods and analysis types are datasets reused?

Hypothesis 1.1: Genomic datasets of the type found in dbGaP will be more likely to be used in combination in meta-analyses, while clinical datasets of the type found in the NIDDK repository will be more likely to be used on their own to answer an original research question.

Research Question 1.2: How closely are the topics for data reuse aligned with the topics for which the data were originally collected?

Hypothesis 1.2: Similarity between original topics and topics of reuse will be lower for genomic data (found in dbGaP) than for clinical data (found in the NIDDK repository).

Research Question 2: What are the demographics of researchers who reuse existing datasets?

Research Question 2.1: Where are requestors located in the world?

Hypothesis 2.1: Requestors will be primarily located in regions with a greater proportion of research institutions, including North America, Europe, and Asia.

Research Question 2.2: Are there patterns in career stage of requestors?

Hypothesis 2.2: A broad range of career stages, from student to full professor (or equivalent) will be represented.

Research Question 3: Are there temporal patterns to dataset requests?

Hypothesis 3: Patterns of requests relative to the original dataset release date will demonstrate a cumulative advantage process, similar to other scientific communication processes such as article citation.

Research Question 4: Are there dataset topics that are more highly requested?

3.1 Research Design

This study utilizes a mixed methods approach to explore the complicated phenomenon of biomedical data reuse, employing both manual techniques for analyzing the qualitative content of data reuse requests and automated analyses that aim to quantify and better understand patterns of requests. A mixed methods design has the benefit of combining qualitative and quantitative methods to provide a more complete picture of a phenomenon, as well as allowing for exploration of multiple related research questions (Bryman, 2006). This chapter describes how this study

combines qualitative content analysis of data reuse requests with quantitative methods, including text mining and bibliometric modeling.

3.1.1 Operationalizing "Reuse"

Quantifying reuse of datasets is challenging, since reuse can take so many different forms. Some forms of reuse are easy to identify, but others leave few traces that can be identified and tracked. An article that makes an explicit citation to a shared dataset and clearly describes its role in the study is an obvious instance of reuse; however, articles frequently do not cite datasets in systematic ways that can be easily and automatically tracked. Even when efforts are made to systematically track and record citations to datasets, dataset requests typically outnumber citations by 75% (Federer, 2018). While using citations as a proxy likely underestimates reuse, using counts of downloads and views as a proxy likely *overestimates* reuse. In open repositories where anyone can download or view a dataset, it cannot be known how or even if the downloader goes on to use the data. Further, because most of these repositories do not collect information about who is viewing or downloading, little can be known about the potential users of the dataset.

One approach that may more accurately reflect reuse is analysis of data use requests. Repositories that contain sensitive human research data cannot make datasets available to freely download because of privacy and consent issues. Instead, researchers must make a formal request for datasets, including a description of the specific purpose for which they are requesting the data and, in most cases, clearance

from their Institutional Review Board (IRB); these requests are then reviewed by a Data Access Committee (DAC) at the repository, a body charged with determining acceptable reuse. Since researchers cannot use the data without submitting a request, and a request cannot be submitted without having a specific intended use, use requests likely provide a reasonably complete representation of data reuse, as well as providing information about how the requestor intends to use the data. In this study, I will draw on these requests as a proxy for reuse.

While they are likely more accurate than citations or download counts, use requests also do not provide an exact measure of data reuse. Just because researchers must have a specific use in mind when they apply for the dataset does not mean that they end up using the data. They may realize once they have the dataset that it is not actually suited for their purpose after all, or they may discover that the data do not support their initial hypothesis and discard the project. Knowing the identity of the requestor also does not mean that the actual data reuser is known; it is possible that someone else, or even whole research groups, are the actual users. For example, a professor might request a dataset on behalf of a student, or an administrator may request data on behalf of an entire team. Despite these limitations, use requests are a useful proxy for data reuse in the context of this study, in that they provide a depth of information about how researchers at least *intend* to reuse datasets. Throughout this dissertation, I will discuss how the limitations of this approach constrain the application and generalizability of the findings, as well as propose how these findings

could be supplemented by future research that draws on other methodologies and data sources.

3.1.2 Sampling and Data Collection

This study considers three repositories administered by various groups within the NIH, all of which require researchers to submit requests to reuse the datasets. While other NIH repositories do exist, these three lend themselves to study because they not only require submission of use requests, but also make most or all of the request contents publicly available. The Database of Genotypes and Phenotypes (dbGaP), administered by the National Center for Biotechnology Information (NCBI), contains human genetic sequence data and associated diseases or characteristics (National Center for Biotechnology Information, 2018). The BioLINCC repository and the NIDDK Central Repository contains biospecimens and datasets arising from research funded by the National Heart, Lung, and Blood Institute (NHLBI) and the National Institute of Diabetes and Digestive and Kidney Diseases' (NIDDK), respectively (National Heart, Lung, and Blood Institute, 2018; National Institute of Diabetes and Digestive and Kidney Diseases, 2018).

Together, these three repositories cover a range of data types, from clinically-focused data (NIDDK and NHLBI) to genomic data (dbGaP), as well as a range of diseases and topics. The data contained within these repositories almost exclusively comes from NIH-funded studies. While individual researchers may submit data, many

of these datasets arise from large efforts that involve research teams or even multi-site consortia.

In addition to use requests, all three of the repositories considered in this study display descriptive metadata about available datasets, including Medical Subject Headings (MeSH) terms that describe the focus of the dataset and narrative descriptions of the original study, which contain information such as the purpose of the original study, data collection methods, characteristics of the original study participants (such as adults, children, healthy volunteers, or individuals with a particular disease), and findings of the original study.

Table 3-1 summarizes the counts of data used in this study (the total datasets requested is greater than total requests, since many individual requests mentioned more than one dataset). NHLBI could not provide identifying information about requestors for privacy reasons; therefore, analyses can be conducted at the dataset and institution levels, but not individual requestor level for NHLBI. While dbGaP's full dataset also includes requests that were rejected, this study considers (and Table 3-1 reflects) only requests that were accepted. Future study on differences between requests that were accepted and those that were rejected may be fruitful, but this study considers data reuse, which of course did not occur in the case of requests that were rejected. Table 3-2 indicates the content included in the use requests by repository.

Table 3-1. Number of datasets, requestors, institutional affiliations, and use requests from each repository and overall.

	dbGaP	NHLBI	NIDDK	All repositories
Datasets	1,014	146	77	1,237
Total requestors	5,260	N/A	253	5,513
Total institutional affiliations	1,230	1,001	195	2,426
Total requests	9,444	1,939	449	11,832
Total datasets requested	104,326	3,864	562	108,752

Table 3-2. Contents of use requests by repository (X indicates the repository contains the item).

	dbGaP	NHLBI	NIDDK
Requestor name	X		X
Requestor institution/affiliation	X	X	X
Dataset(s) requested	X	X	X
Date of request	X	X	X
Reuse summary		X	X
Technical research use statement	X		
Non-technical research use statement	X		

I acquired the data for analysis through a combination of web-scraping from the public sites (that is, writing a script to automatically fetch and parse the data) and requesting the data from the repositories. I requested data on use requests and dataset descriptions from NIDDK and NHLBI, and they provided this information in a set of comma-separated value (CSV) files. dbGaP staff were not able to provide the information I requested due to staffing limitations and time constraints. However, all

the information needed for these analyses is publicly available on the dbGaP website, so I was able to obtain the necessary data from the site. Rather than manually download all the metadata and use requests, I wrote an R script that automatically downloaded the dbGaP use requests and dataset metadata. This web-scraping process was accomplished using the R packages httr (version 1.3.1) and rvest (version 0.3.2) (Wickham, 2016, 2017a).

Once I obtained all the data, I wrote various custom R scripts to clean, organize, and visualize the data to prepare it for coding and analysis, incorporating existing functions from the R tidyverse package (version 1.2.1) (Wickham, 2017b). Except where noted otherwise, all code is written in R version 3.4.1 and run in RStudio version 1.0.143. All code used for data collection, cleaning, and analysis is available at https://github.com/informationista/integrative_paper.

3.2 Data Preparation and Analysis

3.2.1 Research Question 1: For what research objectives are biomedical datasets reused?

Requests to dbGaP and NIDDK included not only general information about who was making the request and what data they were requesting, but the actual text of the request itself. These requests provide an overview of how the requestor intended to reuse the dataset, written with enough detail to enable the repositories' Data Access Committee (or equivalent body) to make a determination about whether the request constituted valid and appropriate reuse. These detailed requests provide a rich corpus from which to draw information about how data are intended to be reused.

Specifically, I consider the type of reuse and the similarity of the reuse to the topic for which the data were originally collected.

I manually coded requests for the type of reuse from a taxonomy drawn from existing literature and validated in my previous research on the use requests in this dataset (Borgman, 2011; Coady et al., 2017; Federer, 2018; Pasquetto et al., 2017). I also inductively added categories as needed for cases that did not fit within the taxonomy. For example, initial coding revealed that some of the use requests asked for data to include in a larger database for general use, which did not fit any of the existing categories. Therefore, I added the category of "infrastructure" to describe this type of reuse. Table 3-3 describes and defines the categories used in this analysis.

Table 3-3. Coding categories and their definitions.

Category	Definition
Original research study	use of a single dataset to answer a new research question, distinct from the specific question for which the data were originally collected
Meta-analysis study	aggregation or integration of the dataset with other datasets to answer a research question or conduct a formal meta-analysis
Statistical methods study	use of one or more datasets to develop or verify new statistical methodology
Software or tool development study	use of one or more datasets to develop, test, or validate a new software product or analysis tool
Validation	use of one or more datasets to validate other findings, such as validating findings from an animal model in human subjects
Comparison or control	use of one or more datasets to validate the investigator's own data, provide comparison, or serve as a control group
Reproducibility or reanalysis study	reanalysis of one or more datasets to answer the same question for which the data were originally collected or to verify the original study's findings
Infrastructure	use of one or more datasets to populate a database or

Category	Definition
	repository for internal or institutional use

Of the 449 unique requests for NIDDK datasets, 17 were missing the executive summary that contained information about how the datasets would be reused. 432 unique requests had executive summaries, an amount that was small enough to permit me to code all the requests. This total population sampling has the benefit of avoiding sampling error and providing a richer understanding of the phenomena of interest (Etikan, Abubakar Musa, & Sunusi Alkassim, 2016; Thygesen & Ersbøll, 2014). However, dbGaP datasets had 9,444 unique requests, too many for me to feasibly manually code. Therefore, for the dbGaP analysis, I randomly selected a subset of 1,500 of the 9,444 requests (15.9%), which provides a confidence interval of +/-1.1 at a 95% confidence level (based on estimation of proportion).

To identify the topic of reuse proposed in a dataset, I used an automated coding method rather than manual coding. Using an automated method has the benefit of applying systematic coding, not affected by human judgment, across the entire dataset. The use of an automated technique also allowed me to include the entire set of both dbGaP and NIDDK requests (9,444 and 432 requests, respectively) in this analysis, since I was not limited by what I could feasibly manually code.

MeSH On Demand is an automated tool that generates a list of Medical Subject Headings (MeSH) terms for each use request, including the specific organ systems, diseases, research techniques, and other topics that describe the content of a text by using the National Library of Medicine's (NLM) Medical Text Indexer (MTI)

(National Library of Medicine, 2018). MTI was originally developed to partly automate indexing of journal articles for inclusion in MEDLINE, the database of biomedical literature maintained by the NLM. Prior to the development of MTI, human indexers manually indexed all articles for MEDLINE; by 2014, MTI was being used in the indexing of over 60% of MEDLINE articles (Mork, Aronson, & Demner-Fushman, 2017).

With advances in technology and the application of machine learning technologies to MTI, its precision and recall have improved since its original development in 2002 (Aronson, Mork, Gay, Humphrey, & Rogers, 2004; Mork, Yepes, & Aronson, 2013). Even so, MTI is not as accurate as a human indexer, so I tested whether it would perform adequately for use in this study by comparing the terms it automatically generated with my own manual coding for ten randomly selected use requests from the repositories considered in this analysis (five from dbGaP and five from the NIDDK repository). For each of the requests, MTI assigned more terms than I did (mean 9.7 terms per use request for MTI compared to a mean of 4.7 for me). My own indexing focused only on terms such as diseases, conditions, and organ systems, which are the categories of terms that are used to describe the original datasets, while MTI also picked up on concepts such as analytical methods and study populations. Considering only disease, condition, and organ system terms, the MTI terms and my own matched in all ten cases. Given that MTI's sensitivity (in other words, its ability to identify all relevant terms) is similar to my human indexing, its

lack of specificity (that is, its tendency to identify some irrelevant terms) does not present a problem for this study.

To help improve the accuracy of the MTI indexing, I also removed high-level terms related to study populations, such as Male, Female, Child, and Adult. Leaving extraneous terms in would not significantly affect the outcome of the analysis; as will be discussed, the algorithm that calculates similarity considers terms that come from two separate "branches" of the MeSH tree hierarchy to be entirely unrelated. Since the MeSH terms assigned to datasets almost exclusively covered diseases and organ systems, which are on separate branches of the MeSH tree from study population terms, the algorithm would consider these terms unrelated to the dataset terms, which would make them irrelevant to this analysis, since the similarity score is based on the set of most similar pair of terms. Leaving unrelated or extraneous terms in the MTI-produced term lists would therefore have no impact on the outcome of the analysis, but removing them did improve the efficiency of an already computationally intensive analysis, so I removed them.

Once MeSH terms were assigned for each request, these terms could be compared to the MeSH terms assigned by the repositories to the corresponding dataset in order to determine how closely the proposed reuse matches the original reason for which the dataset was collected. This comparison is based on a technique called semantic similarity, which employs ontologies to calculate the relatedness of a set of terms (Pesquita, Faria, Falcão, Lord, & Couto, 2009). MeSH's tree structure makes it possible to calculate semantic similarity between terms based on their

relative positions in the hierarchy, where 0 means two terms are completely unrelated and 1 means they are identical (Gan, Dou, & Jiang, 2013; Garla & Brandt, 2012; Zhou et al., 2015).

Figure 3-1 demonstrates the concept of semantic similarity in a small portion of the MeSH tree structure. Considering the term Heart Diseases, some of the terms in the tree are similar conceptually (for example, Vascular Diseases also affect the cardiovascular system), while others are completely unrelated (for example, Informatics is on a totally separate branch of the MeSH tree and has no conceptual relationship to Heart Disease). Figure 3-1 shows the semantic similarity score (SSS) for each term to the index term of Heart Diseases.

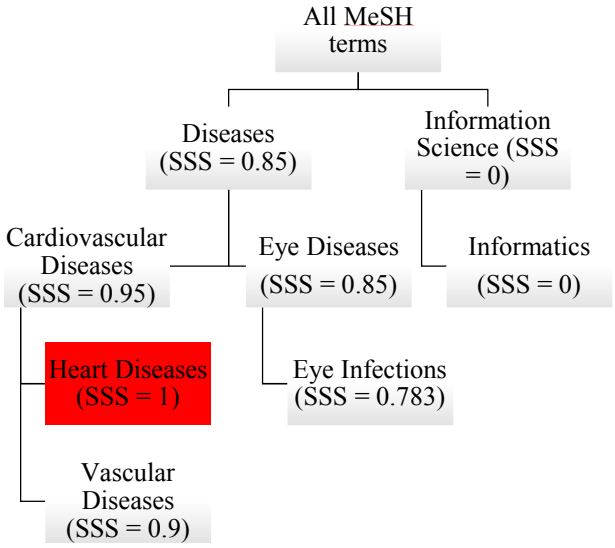

Figure 3-1. MeSH tree sample demonstrating semantic similarity. The number following each term is its semantic similarity score (SSS) to the index term of "Heart Diseases."

I calculated semantic similarity using the shortest path algorithm in the R package MeSHSim (version 1.2.0; requires R version 3.2.1) (Zhou & Shui, 2015). I

tested each of the nine algorithms that are implemented in this R package; all performed similarly in terms of how they relatively ranked similarity of terms, but the shortest path algorithm has the benefit of being on a 0 – 1 scale that enables straightforward interpretation of similarity (or lack thereof). Both use requests and datasets can be tagged with multiple MeSH terms; the MeSHSim package returns results in the form of a matrix of similarities for all terms, as shown in the example in Table 3-4. I recorded the highest semantic similarity value for each use request/dataset pair (for example, in the case of the terms in Table 3-4, I would record the value 0.86, since the terms Lung and Cardiovascular System are most similar). Since the datasets and requests are both described by multiple terms, it is likely that many of the term pairs in the matrix will be 0, even if the dataset and request also share a term that is an exact match. For that reason, the use of the maximum rather than the mean score provides a better understanding of the similarity between the dataset and the request.

Table 3-4 An example matrix of semantic similarity scores between two sets of terms.

		Dataset terms			
		Ankle Brachial Index	Cardiovascular System	Intermittent Claudication	Peripheral Vascular Diseases
Request terms	Lung	0	0.86	0	0
	Smoking	0	0	0	0
	Global Health	0	0	0	0

	Cohort Studies	0.24	0	0	0
	Biological Markers	0	0	0	0
	Pulmonary Disease, Chronic Obstructive	0	0	0.62	0.65

Together, the analyses of manually-coded reuse types and machine-coded topics provide insight into the uses for which data are being requested. For example, are most datasets being used in the context of the same topic for which they were originally collected, as measured by semantic similarity? Are multiple datasets being combined to derive additional findings that would not have been possible using a single dataset on its own? Are genomic datasets of the type found in dbGaP reused in different contexts or ways than clinical datasets of the type found in NIDDK? These findings contribute to a clearer view of biomedical data reuse that will contribute to understanding the impacts of shared datasets. Given the concern that many researchers have about others "scooping" their work if they share their data, the answers to these questions may also have implications for researchers' attitudes toward sharing.

3.2.2 Research Question 2: What are the demographics of researchers who reuse existing datasets?

To better understand the types of requestors who are reusing data, I manually coded the use requests with demographic information about the requestor. First, for

each unique institution, I recorded the latitude and longitude of the institution's city to determine where requests are originating. The latitude and longitude enabled me to use the data with R packages that rely on geocoded data for visualization, both at the international level and at the state level within the United States. The institution name was available for all 9,444 dbGaP requests, all 1,939 of the NHLBI requests, and 255 of the 449 NIDDK requests (57%). The large number of missing institutions in the NIDDK datasets is due to differences in the repository's systems prior to September 2013 that resulted in some data about requests being unavailable; therefore, this analysis reflects only the most recent six years of use.

Raw counts of requests would not provide useful insight into which countries were making the most reuse of shared datasets, since research activities are not evenly distributed around the world. For example, it would be reasonable that more requests would come from the United States (a large country with a sizeable research enterprise) than say, Liechtenstein (one of the smallest countries in the world). Therefore, rather than use raw counts, I compared the number of requests coming from a geographic region to its research presence. Research presence is difficult to quantify, since research is conducted within many different organizations, including academic institutions, government agencies, non-profit organizations, and private research corporations, to name a few. For international-level comparisons, I used number of universities as a proxy for research presence. For state-level comparisons, I used NIH funding received within each state in Fiscal Year 2018 (the most recent year for which complete funding data are available). This state-level proxy likely

provides a more accurate representation of research presence, since NIH funds are awarded not only to universities, but to other types of research institutions.

To compare a country's (or state's) research presence to the number of use requests its researchers make to each repository, I calculated the relative difference in composition (RDC). RDC is a measure of how over- or underrepresented a group is within a specific context compared to the composition of the entire population. For example, RDC has been used to measure underrepresentation of racial groups in gifted and talented education programs compared to their total presence in a school overall; a group that makes up 50% of the students in the whole school, but only 25% of the students in the gifted and talented program is underrepresented (Ford, 2014). I calculated RDC for countries and for states within the United States to determine whether certain geographic regions are making more requests than might be expected based on their research presence. I did this analysis for each repository individually to determine whether there was variation in where requests were concentrated for the different repositories.

To better understand who is reusing data, I also coded each unique request with the requestor's career stage at the time of the request. To determine career stage, I located web resources that documented requestors' career, such as LinkedIn, CVs, biosketches, and web pages. Where I could not definitively determine a requestor's career stage using available online materials, I coded the career status as unknown. Because a single requestor may have made multiple requests across his or her career, I recorded the career stage for each unique request. For example, a requestor may

have been an assistant professor when she made her first request in 2013, but she had received tenure and was an associate professor by the time she made her next request in 2016. I converted non-United States job titles to their United States equivalent to allow for comparison across countries. For example, in many commonwealth countries such as the United Kingdom and Australia, the term "lecturer" is the equivalent of assistant professor in the United States (Wikipedia, 2018).

Because NHLBI did not provide the names of individual requestors, I limited this analysis to the dbGaP and NIDDK requests. Of the 449 unique NIDDK requests, 286 included the requestor's name (64%). As with institution name, the requestor names were missing from the oldest requests (in the case of requestor name, those made before December 2012). The 9,444 requests to dbGaP came from 5,260 unique requestors. As with coding for reuse type, locating career status information for so many requestors was not feasible, so I coded a subset of 1,500 of the 9,444 requests (15.9%), which provides a confidence interval of +/-1.1 at a 95% confidence level (based on estimation of proportion).

As with the distribution of research across different geographic areas, the distribution of researchers across career stages is not totally even. More requests may come from assistant professors simply because more researchers are at this career stage, and not because they are actually making more requests than researchers at other stages. Therefore, I took the same approach of calculating relative difference in composition between the proportion of individuals at a career stage overall and the number of requests coming from individuals at this career stage. For non-academic

career stages, determining the number of individuals in a given career stage, such as "senior scientist" or "executive" would be nearly impossible, since these individuals are employed in so many different types of institutions. However, for academic requestors, this analysis is possible, since the National Center for Education Statistics tracks counts of full-time faculty in US degree-granting postsecondary institutions (National Center for Education Statistics, 2017).

One unavoidable limitation of this approach is that the person who requested the data may not actually be the person who *used* the data. For example, a junior lab member may request data on behalf of his or her principal investigator, or a professor may request data on behalf of a student. Future survey research of dataset requestors could help elucidate the extent to which the data requestor and the data reuser differ.

3.2.3 Research Question 3: Are there temporal patterns to dataset requests?

The repositories in this study contain many years' worth of datasets, some dating back to the early 2000s, and records of requests dating back almost as long. With many years' worth of request data available, it is possible to track the dynamics of requests over a dataset's lifetime to better understand when datasets are most requested. Further, understanding temporal patterns to dataset requests could make it possible to predict early in a dataset's life how much use it would receive in the long-term, which could be useful in making curation and preservation decisions. Knowing how long a dataset remains useful could also influence preservation decisions – if

datasets are generally no longer requested once they reach a certain age, it may be reasonable to discard them.

This inquiry into requests to datasets over time is similar to the study of citation dynamics within bibliometrics, which considers the numbers of citations an article receives over time. Part of this exploration involves mapping "citation bursts," or the time it takes for articles in a given field to reach their peak annual citation before citations begin to decline (Eom & Fortunato, 2011). The literature also contains explorations of unique or unusual citation dynamics, such as descriptions of the dynamics of "sleeping beauties" (articles that receive few citations for many years and then suddenly attract significant attention) and "flashes in the pan" (articles that receive a great deal of initial attention, which quickly dies down) (Li, 2014; van Raan, 2004). These explorations provide a basis upon which to begin to explore temporal patterns of dataset requests.

As has been previously discussed, NIDDK's move to a different system in September 2013 means that the year of release for datasets prior to that date is unknown. Removing all datasets from before September 2013 left too few datasets for this analysis. Therefore, for this analysis I used dbGaP, which contains 982 datasets with a total of 100,115 requests, and NHLBI, which contains 143 datasets with a total of 3,860 requests. For each dataset, I aggregated the number of requests it had received each year. I also calculated the dataset's age at the time of request, enabling comparison across datasets at the same age, regardless of when they were released. If dataset requests are a cumulative advantage process, with success

breeding success, then datasets that are older are likely to receive more requests in a given calendar year than those that are younger. For example, consider a dataset released in 2010 and one in released in 2016. The older dataset has had an additional six years to accrue advantage, so if we compare the number of requests each received in calendar year 2017, it is likely that the 2010 dataset would receive more than the 2016 dataset. However, considering how many requests each received in the first year after they were released provides a more meaningful basis for comparison.

Once the number of datasets requested per year of a dataset's life was calculated, I divided the data within each repository into groups. First, I divided the datasets into tiers based on their percentile ranking of total requests over time, that is, the top 10% most requested, the next 10% most requested, and so on. To better control for age of dataset, I also calculated the mean percentile ranking over the course of a dataset's life. For example, if it was in the 20th percentile of first year requests, the 30th percentile of second year requests, and the 40th percentile of third year requests, its mean percentile ranking is 30th percentile. I divided the mean percentile rankings into quartiles. I then plotted both the overall request deciles and the mean request quartiles to visualize the pattern of requests for datasets of varying levels of attention based on requests.

In addition to understanding patterns of requests over time, I also aimed to determine whether the number of requests a dataset received early in its life was predictive of how many requests it would receive over the long run. That is, does a dataset that receives many requests in its first year likely to go on to receive more

requests than a dataset that is less requested soon after its release? I tested this by fitting three regression models to the dbGaP and NHLBI dataset, looking at the relationship between total requests and first-year requests only; first- and second-year requests; and first-, second-, and third-year requests, controlling for year of release for all three models. These models provide an understanding of the extent to which requests in the first three years of a dataset's life can be used to potentially predict the number of requests it will go on to receive.

Because dynamics and temporal patterns of dataset requests have not yet been studied, my primary aim here was to determine whether in fact patterns do indeed exist, and if so, the general dynamics of requests over time. This study provides an initial view of the temporal patterns within dataset requests, that can be expanded based on request dynamics. This analysis also demonstrates the extent to which dataset requests can be considered a cumulative advantage process.

3.2.4 Research Question 4: Are there dataset topics that are more highly requested?

The time-based methods described above provide insight into patterns of how datasets are requested over time and whether cumulative advantage processes and attention decay effects influence how many requests datasets receive. However, these models likely do not fully account for the reasons why some datasets are more highly requested than others. Previous studies have explored researchers' decision-making processes related to choice of and satisfaction with datasets, but the factors identified in these studies are subjective and would be difficult to measure in the context of this

study. For example, opinions about dataset credibility would likely differ significantly among dataset requestors, so it would be difficult to develop a method to quantify credibility as a factor. Reputation of the data creator is also a factor in data reuse; even if there were an objective measure of reputation, many of these datasets have been collected by large, multi-site consortia with many individuals involved, and some of the datasets do not list who originally collected the data at all. In the absence of robust and reliable methods for quantifying these subjective measures, it is necessary to look to the datasets themselves to understand why some are more highly requested.

The repositories considered in this study do include some basic metadata about the dataset, such as the number of subjects in the dataset and the dates of data collection. However, this metadata is sparse and provides little useful insight into the content of the dataset itself. In addition, the content of the metadata differs across the three repositories, making it challenging to identify patterns that would hold for biomedical data reuse broadly, rather than being specific to an individual repository.

More useful than this basic metadata is the narrative description of the dataset, which can be meaningfully explored using text-mining methods. At its most basic level, text mining is useful in understanding the contents of a document by identifying the terms that are most central based on frequency (Hotho, Andreas, & Paaß, 2005). This simple approach considers a document as a "bag of words," simply counting the number of times a given word appears without consideration of its context within the text (Y. Zhang, Jin, & Zhou, 2010). More advanced topic modeling techniques make

it possible to identify complex latent topics in a text by counting words in their broader context, such as considering *n*-grams (a set of *n* words appearing together in sequence), sentences, or paragraphs (Blei, Ng, & Jordan, 2003).

These topic modeling techniques are considered "unsupervised," in that the algorithm simply identifies patterns of words within a corpus that frequently appear together in texts, and it is up to a human subject matter expert to determine the topic it describes. For example, the algorithm might determine that the terms "myocardial infarction," "hypertension," and "cardiac output" form a topic in a corpus; a human interpreter would then be able to determine that texts containing this topic could be described as being about "cardiovascular disease."

Text mining is especially useful in this analysis because it allows for the detection of patterns in the data even when potentially important features are not known in advance and has the benefit of being able to account for a wide range of features that are not captured in the metadata. For example, since data descriptions include information such as the specific brand of the sequencing machine and the study methodologies, these methods will be able to take into account whether these features are characteristic of reuse.

Text mining techniques also are a practical method here because they have been demonstrated to be useful in various bibliometric applications, which, as has been discussed, is similar to the type of inquiry being conducted here. For example, topic modeling techniques have been used to successfully identify high impact articles, with significant correlation to article citation counts (Gerrish & Blei, 2010;

Mann, Mimno, & McCallum, 2006). Text mining has also been used to detect similarities between patent documents and scientific articles (Magerman, van Looy, & Song, 2010); while I did not use that technique in this study, this approach could have potential future applications for detecting similarities between dataset descriptions and the associated reuse requests.

The narrative study descriptions from each of the three repositories formed the corpus for text mining, specifically using a topic modeling approach. This analysis includes the descriptions of 1,150 datasets from dbGaP, 166 datasets from NHLBI, and 140 datasets from NIDDK. I wrote a script that retrieved dataset descriptions from the webpages of each of the datasets' web pages, then prepared the texts using standard text mining pre-processing techniques incorporated in the R text mining package tm (version 0.7-3) (Meyer, Hornik, & Feinerer, 2008), including converting all text to lowercase (since R is case-sensitive); removing common English language stopwords such as "the" and "and"; stemming, which converts various forms of a word to their common root (for example, "genetic," "genetically," and "genetics" would all be collapsed to "genetic"); and trimming of white space and special characters. The dataset descriptions, particularly from the same repositories, are all somewhat homogenous in terms of certain scientific words that would not be contained in the English language stopword list, but that would not be informative about the content of the description, such as "study" and "subject." Therefore, I also removed a custom set of stopwords that appeared almost universally in the

descriptions and provided no useful context about the topic of the dataset; this list is in Appendix B.

Once the texts were prepared, I proceeded to develop topic models for each repository using latent Dirichlet allocation (LDA) implemented in the R package topicmodels (version 0.2-7). To understand the LDA model, consider a set of documents from a corpus. Most documents do not have a single topic, but several; for example, a description of a dataset in dbGaP has the topic genetics, as well as the topic of whatever disease or condition it is studying. The topic genetics, in turn, has a number of words that are associated with it, such as "genetic," "sequence," and "genome." LDA fits a mathematical model to "[find] the mixture of words that is associated with each topic, while also determining the mixture of topics that describes each document" (Silge & Robinson, 2018, para. 5). The topicmodels package will generate a list of the terms most highly associated with each topic, as well as calculating the probability that a term is predictive of a given topic. For example, the term "carcinoma" would have a higher probability of being associated with the topic of cancer than the topic of cardiovascular disease. The word "topic" should be understood broadly here, not just to refer to the disciplinary focus of a dataset, but to also potentially draw on other concepts contained in the descriptions, such as study type or characteristics of subjects.

The application of the LDA model does require a fair amount of judgment on the part of the human programmer. For example, the choice must be made whether to use individual words as the "token" or unit of analysis (the bag of words approach) or

group words into *n*-grams with their nearest neighbors. For example, in a simple text with a basic vocabulary, the bag of words approach may be effective, but more technical texts might use many multi-word phrases, which would not be reflected if using simple counts of single words. Therefore, achieving meaningful results requires experimenting with using single words, bigrams (word pairs), or trigrams (word triplets). In addition, the human must determine the number of topic groups into which to divide the corpus. While there are some statistical methods that can aid in identifying the optimal number of topic groups, achieving meaningful topics largely relies on human judgment. The process of determining the number of topics is iterative, starting with the predicted optimal number of groups and experimenting with the varying numbers until the most meaningful categories appear. In addition, it is up to the human to identify what the topics actually describe. The topicmodels package simply returns a set of numbered topics and the words most highly associated with them; based on the mixtures of terms associated with the topic, I applied my subject matter knowledge to determine what the topic describes.

Once the datasets were organized into topics, I determined which topics were most requested based on the number of use requests to datasets in each topic. I looked not only at overall counts, but counts by year, to determine whether the most popular topics changed over time. A problem here is that the datasets are not divided evenly among the topics. For example, in Figure 3-2, Topic A contains 4 datasets while Topic B contains only half as many. Topic A would be reasonably expected to have more requests than Topic B, not necessarily because it's more popular, but because it

contains more datasets to receive requests. The fact that Topic B has actually received more requests than Topic A despite having fewer datasets must also be accounted for; not only does it have more requests, but it has them despite having half as many datasets to be requested.

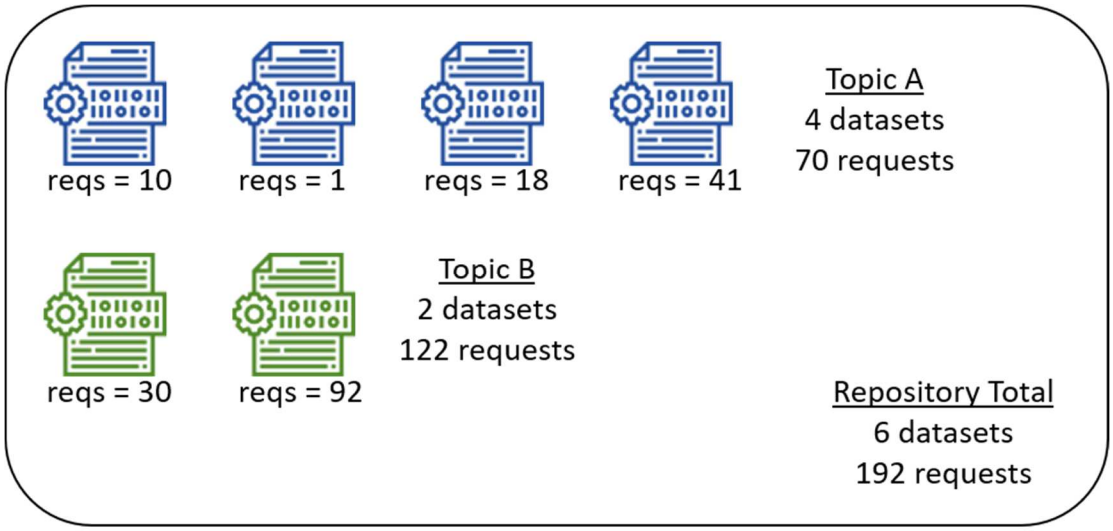

Figure 3-2. Demonstration of analyzing topics and requests.

To solve the problem of comparing topics of uneven size, I compared the proportion of datasets in a topic to total datasets in the repository, to the proportion of requests received by the topic to total requests received by the repository. For example, Topic A contains 4 datasets of the 6 datasets total (0.67) and 70 requests of the 192 requests total (0.36). That is, it contains 67% of the total datasets but has received only 36% of the total requests. By comparison, Topic B only contains 33% of the requests but received 64% of the requests. This analysis makes it possible to compare topics' requests even when the datasets are unevenly distributed among them, to determine which topics are most highly requested.

3.3 Limitations

It is important to note that the scope of this study limits its generalizability not only beyond biomedical data, but also beyond the three repositories considered here. Analyzing each repository separately from the others makes it possible to gain insight into the extent to which biomedical repositories differ from each other, such as whether genomic datasets are reused differently from clinical datasets. Still, caution should be used in generalizing results, and further research should examine whether the findings of this study hold for other repositories, data types, and disciplines.

The repositories considered here are also somewhat unique in that they are restricted access repositories. Because of the limitations I have described, it is difficult or even impossible to know who is using data from truly open repositories and in what ways. Counts of dataset views and downloads provide limited insight into the deeper questions about dataset reuse considered here. At present, use requests are one of the few robust ways to operationalize data reuse, so the limitations associated with these findings are difficult to avoid. However, efforts currently underway in the scientific community to standardize data citation will likely enable better automated tracking of data reuse over time, including data from both restricted access and fully open repositories. As data citation standards mature, future research may be able to address questions about differences in reuse of different types of data and repositories.

Chapter 4: Findings About Requests and Requestors

This chapter presents the findings of the two research questions that focus on questions about requests and requestors, or the who, where, and why of biomedical data reuse – who is reusing biomedical data, from where in the world do requests come, and why are datasets reused. Specifically, the research questions and hypotheses considered here are:

Research Question 1: What are the purposes and characteristics of biomedical research reuse?

Research Question 1.1: For what methods and analysis types are datasets reused?

Hypothesis 1.1: Genomic datasets of the type found in dbGaP will be more likely to be used in combination in meta-analyses, while clinical datasets of the type found in the NIDDK repository will be more likely to be used on their own to answer an original research question.

Research Question 1.2: How closely are the topics for data reuse aligned with the topics for which the data were originally collected?

Hypothesis 1.2: Similarity between original topics and topics of reuse will be lower for genomic data (found in dbGaP) than for clinical data (found in the NIDDK repository).

Research Question 2: What are the demographics of researchers who reuse existing datasets?

Research Question 2.1: Where are requestors located in the world?

Hypothesis 2.1: Requestors will be primarily located in regions with a greater proportion of research institutions, including North America, Europe, and Asia.

Research Question 2.2: Are there patterns in career stage of requestors?

Hypothesis 2.2 A broad range of career stages, from student to full professor (or equivalent) will be represented.

4.1 Research Question 1: For what research objectives are biomedical datasets reused?

Biomedical research is a large umbrella that encompasses many different research methodologies on a range of topics, from efforts aimed at understanding the very building blocks of life to specific trials on the efficacy of various types of therapies. The types of data that comprise biomedical research data are similarly diverse – as are their potential applications. Even when datasets seem very specific in their scope and application, the potential often exists for researchers to reuse data in new and sometimes unexpected ways. In fact, as data science methodologies advance, biomedical research data has potential for researchers who might not even be considered biomedical researchers, such as computer scientists who need test data to develop and validate new algorithms or statisticians who can use existing data to pioneer new statistical approaches.

Here, I aim to better understand how researchers are making use of data available through the dbGaP and NIDDK repositories by examining the descriptions that are submitted as part of a potential reuser's request to access the data. These descriptions, intended for evaluation by repository staff responsible for determining whether the use is appropriate, contain details about the specific research questions researchers intend to explore with the dataset. In this section, I use a combination of qualitative analysis and computational indexing methods to understand the types of research conducted using these datasets, as well as the topics of reuse, including how similar (or different) they are from the original data use.

This analysis draws on data from two repositories; NIDDK provided me a spreadsheet containing details of use requests, including the proposed use, and I wrote an R script to retrieve requests from the dbGaP website. The NIDDK requests cover the period between 2005 and 2018, while dbGaP includes 2007 to 2018. NHLBI does not make their full use requests public; they provided me with summary information about requests, but they did not share identifying information about requestors or the text of proposed uses. Therefore, this analysis does not include NHLBI requests.

4.1.1 Research Question 1.1: For what methods and analysis types are datasets reused?

To determine the purpose for which researchers intend to use requested datasets, I analyzed the descriptions of reuse that are included in the request submission. I hypothesized that the types of reuse described for datasets in dbGaP, which contains primarily genetic data, would differ from those in NIDDK, which

contains primarily clinical data. Given that studies using genetic data typically require a large number of subjects to achieve adequate statistical power (Hong & Park, 2012), I expected that dbGaP would have more requests to use datasets in meta-analyses (that is, in combination with other data). On the other hand, clinical data of the type found in the NIDDK repository can be difficult to combine with other datasets because of nuances of how individual researchers or teams collect the data, so I would expect that these datasets would more likely be used on their own to answer an original research question.

After reading the proposed use for each request, I classified the request according to the type of reuse. The categories were based on review of the relevant literature, with the addition of new categories when needed for use requests proposing an activity not covered by an existing category, and include the eight types described in Table 4-1.

Table 4-1. Coding categories and their definitions.

Category	Definition
Original research study	use of a single dataset to answer a new research question, distinct from the specific question for which the data were originally collected
Meta-analysis study	aggregation or integration of the dataset with other datasets to answer a research question or conduct a formal meta-analysis
Statistical methods study	use of one or more datasets to develop or verify new statistical methodology
Software or tool development study	use of one or more datasets to develop, test, or validate a new software product or analysis tool
Validation	use of one or more datasets to validate other findings, such as validating findings from an animal model in human subjects
Comparison or control	use of one or more datasets to validate the

Category	Definition
	investigator's own data, provide comparison, or serve as a control group
Reproducibility or reanalysis study	reanalysis of one or more datasets to answer the same question for which the data were originally collected or to verify the original study's findings
Infrastructure	use of one or more datasets to populate a database or repository for internal or institutional use

Each of the requests may ask for more than one dataset, and I report the findings at the dataset rather than the request level. For example, if I code a request as being a meta-analysis, and it asks for 200 datasets, 200 instances of meta-analysis are added to the tally. This treats each dataset request as its own unit; even though a requestor may use the same request text for more than one dataset, each dataset's request should still be counted. Appendix A provides examples of use requests in each category from dbGaP and NIDDK.

The determination of how to categorize each dataset was based on a number of factors, including the number of datasets included in the request (more than one requested dataset would suggest a meta-analysis) and the inclusion of phrases that explicitly named a reuse type (e.g. "we propose to *validate* findings from our own colorectal cancer studies" or "our goal here is to perform a *meta-analysis* of densely sequenced genomes" – emphasis added) or keywords that likewise identified a reuse type (e.g. "we develop a Bayesian hierarchical model," with Bayesian referring to a *statistical* approach or "we are currently evaluating the performance of our mutation detection pipeline," where a pipeline refers to a series of *software* tools used in sequence to conduct a specific analysis).

I also drew on my own extensive experience with biomedical research, including nearly ten years working in biomedical libraries, the last six of which were spent working closely with researchers at the National Institutes of Health and National Library of Medicine. In that capacity, I have served as a consultant to and collaborator with biomedical researchers, used my expertise in data science as a team member in "hackathons" aimed at using some of these same types of data to answer biomedical research questions, and developed and delivered training for other biomedical librarians interested in learning more about these skills. These experiences have given me a depth of understanding of research techniques and a familiarity with the vocabulary of the science described within these requests.

I also validated my coding by comparing my codes to those of two outside coders for a random subset of twenty requests (ten from each repository). Both coders have experience working with research of the type described in the use requests: one is an academic biomedical librarian who consults with researchers on issues related to biomedical data and computational reproducibility, and the other is an NIH fellow in data science and open science policy, who holds a doctoral degree in computational biology. Their mean percent agreement with my codes was 72.5% (70% and 75%), which is considered Substantial agreement on Landis and Koch's scale for Strength of Agreement (Landis & Koch, 1977). Most of the variability between their coding and mine was due to their use of the "comparison or control" code when I used "meta-analysis" or vice versa. These two types of reuse are similar, since they both refer to combining data.

The set of NIDDK requests included 416 requests from 252 unique requestors, requesting a total of 561 datasets. Each request asked for a mean of 1.3 datasets, with a minimum of 1 and a maximum of 10. For the dbGaP analysis, I randomly selected a subset of 1,500 of the 9,444 requests (15.9%), which provides a confidence interval of +/-1.1 at a 95% confidence level (based on estimation of proportion). This set came from 1,069 unique requestors and included requests for a total of 20,179 datasets. Each request asked for a mean of 13.5 datasets, with a minimum of 1 and a maximum of 398.

Table 4-2 shows the number of requests in each reuse category and the percent of overall requests for requests to dbGaP and NIDDK.

Table 4-2. Counts and percentages of requests describing various types of reuse for NIDDK and dbGaP datasets.

Reuse type	**dbGaP Requests**		**NIDDK requests**	
	N	**%**	**N**	**%**
Original research	460	2.3%	282	50.27%
Meta-analysis	14,619	72.4%	139	24.78%
Comparison	858	4.3%	2	0.36%
Validation	221	1.2%	14	2.5%
Statistics	2,242	11.1%	84	15.0%
Software	1,097	5.4%	14	2.5%
Infrastructure	644	3.2%	0	0%
Re-analysis	11	0.05%	2	0.36%
Reuse type not specified	2	0.01%	24	4.28%

Although some types of reuse are uniformly low for both dbGaP and NIDDK datasets, the most common ways that they are reused are very different from each other. A chi-squared test of independence confirms that the distributions of reuse between dbGaP and NIDDK are significantly different ($\chi 2 = 4547$, df = 8, $p < 0.01$).

As hypothesized, original research is the most common reuse type for NIDDK datasets, but it is actually the fourth *least* common reuse type for dbGaP datasets. On the other hand, nearly three-quarters of dbGaP datasets are requested for use in a meta-analysis; while meta-analysis is still a significant category for NIDDK data reuse, it is much less common than for dbGaP. The greater frequency of meta-analyses in dbGaP and original research studies in NIDDK is also reflected in the very different number of datasets per request for these two repositories: on average, NIDDK requests ask for just 1.3 datasets to dbGaP's mean of 13.5. A Welch unpaired two-sample t-test shows that the means of datasets per request for dbGaP and NIDDK are significantly different ($t = 11.5$, $df = 1504$, $p < 0.001$).

These variations are likely due to the differences in the types of data that dbGaP and NIDDK house. Genome-wide association studies, a common use of the dbGaP datasets, require a much larger sample size to achieve adequate statistical power than do clinical studies, and therefore several datasets may need to be pooled in order to have enough subjects for a study (Hong & Park, 2012). On the other hand, many of the NIDDK are clinical datasets, which are often difficult to combine using meta-analytic techniques because different research teams collecting the original datasets often use their own unique ways of recording variables.

For example, many of the studies in NIDDK ask participants about their alcohol consumption habits, but they do so in ways that make it difficult to compare across studies. One such study, the Diabetes Prevention Program Outcomes Study (DPPOS) queries in specific detail, asking participants to recall how many "12 ounce

bottles of beer," "4 ounce glass of wine," and "1.5 ounce shots of hard liquor or mixed drinks" they had consumed in the past seven days (Diabetes Prevention Program Outcomes Study, 2016). Another, the Nonalcoholic Fatty Liver Disease (NAFLD) study, simply asks participants how many drinks they have on a typical day (Nonalcoholic Fatty Liver Disease (NAFLD) Adult Database, 2016). It is difficult to know if responses to these questions yield truly comparable results. Perhaps an NAFLD participant is in the habit of going to the pub for a pint of beer (16 ounces) every evening, and without this more specific guidance of "12 ounces," will likely count each of these as one drink. When this NAFLD participant responds he has seven beers a week, he has consumed 112 ounces of beer, or 30% more than a DPPOS respondent who says she consumes seven 12-ounce bottles of beer a week (or 84 ounces). These two studies also differ on how they define binge drinking, with the DPPOS asking about how often the participant has had seven or more drinks in 24 hours, whereas NAFLD asks about how often the participant has had six or more drinks on one occasion. Even such seemingly inconsequential differences – six versus seven drinks, "on one occasion" versus in 24 hours – mean that different information is being elicited from participants. With many of these clinical studies having hundreds or even thousands of variables, these small differences can add up to significant challenges that prevent datasets from being combined for meta-analytic purposes.

4.1.2 Research Question 1.2: How closely are the topics for data reuse aligned with the topics for which the data were originally collected?

This analysis aims to quantify similarity between the original subject focus of shared datasets and the focus of the research for which requestors hope to reuse them. I hypothesize that the differences in genomic versus clinical research discussed above will also lead to differences in the similarity of reuse to original data purpose between dbGaP and NIDDK. Given the broader applications of dbGaP data compared to the relatively specific applicability of NIDDK's clinical datasets, I expect greater similarity between NIDDK datasets and their topics of reuse than for dbGaP datasets and their topics of reuse.

Medical Subject Heading (MeSH) terms provide a means by which to compute an objective measure of similarity between original use and reuse. These terms are used to describe medical literature consistently as well as to understand relationships between terms. Because MeSH terms are arranged in a hierarchical fashion in a tree structure, it is possible to calculate a measure of similarity between two terms, known as semantic similarity. Terms closer to each other in the hierarchy will have a high semantic similarity score, whereas terms that are far from each other on the tree will have a lower semantic similarity score. Exactly identical terms have a semantic similarity score of 1, whereas a semantic similarity score of 0 indicates that the two terms are not in any way topically related (since they are on totally different top-level branches in the 16-branch MeSH tree). Thus, comparing MeSH terms that are assigned to a dataset with MeSH terms assigned to a request for that data allows

for a quantitative measure of similarity between the proposed reuse and the original dataset's purpose.

Conveniently, datasets from dbGaP and NIDDK were classified by the repository with one or more MeSH terms. To determine MeSH terms for the requests, I used the MeSH On Demand tool, which utilizes the National Library of Medicine's (NLM) Medical Text Indexer (MTI) to assign terms to a provided text. Given a reuse request description, the MeSH On Demand tool returns a list of relevant MeSH terms. I removed very general terms, such as "Human" and "Adult" from the list of returned MeSH terms, since these provided little useful context.

Once the terms had been assigned, I wrote an R script that would join the set of MeSH terms for a request with the set of MeSH terms for all the datasets included in the request. Since most datasets and requests had more than one MeSH term, the script calculated a semantic similarity score for each request/dataset term pair and recorded the highest score. Table 4-3 shows an example of a request/dataset pair from dbGaP with their terms and the semantic similarity score for each term. Most of the term pairs have a semantic similarity score of 0, since they are on totally different top-level branches of the MeSH tree. Others have a small score because they are on the same branch, but far apart from each other. For example, Ankle Brachial Index and Cohort Studies are both on the top-level branch Analytical, Diagnostic, and Therapeutic Techniques, and Equipment. However, moving down the tree, they are far down on very distant branches from each other. On the other hand, Pulmonary Disease, Chronic Obstructive, is much closer to Intermittent Claudication and

Peripheral Vascular Diseases on the Diseases top-level branch. Cardiovascular System and Lung are the closest to each other, just one level apart on the Anatomy branch. Because many of the terms are unrelated, recording the maximum score provides the best comparison of the similarity between the request and the dataset; for example, in this case, the mean would only be 0.03, compared to the maximum score of 0.86.

Table 4-3. Example semantic similarity scoring.

		Dataset terms			
		Ankle Brachial Index	Cardiovascular System	Intermittent Claudication	Peripheral Vascular Diseases
Request terms	Lung	0	0.86	0	0
	Smoking	0	0	0	0
	Global Health	0	0	0	0
	Cohort Studies	0.24	0	0	0
	Biological Markers	0	0	0	0
	Pulmonary Disease, Chronic Obstructive	0	0	0.62	0.65

Semantic similarity scores were calculated for each request/dataset pair in dbGaP and NIDDK; NHLBI was not included in this analysis because they did not provide me the text of use requests. The dbGaP dataset included 9,348 unique

requests for 986 unique datasets, for a total of 92,523 request/dataset pairs. The NIDDK dataset included 544 unique requests for 65 unique datasets, for a total of 539 request/dataset pairs. Figure 4-1 shows the distribution of maximum semantic similarity scores for dbGaP and NIDDK request/dataset pairs. The top part of the chart shows density at each score (i.e. the proportion of how many request/dataset pairs have that score). The horizontal boxplot below shows the distribution of maximum semantic similarity scores. Each of the points overlaying the boxplot corresponds to a single request/dataset pair at that score. Table 4-4 provides summary statistics.

Table 4-4. Summary statistics of semantic similarity scores for dbGaP and NIDDK request/dataset pairs.

	Mean score	**Number of pairs with score = 0**	**Number of pairs with score = 1**
dbGaP	0.56	28,804 (31.1%)	18,347 (19%)
NIDDK	0.78	85 (15.8%)	297 (55.1%)

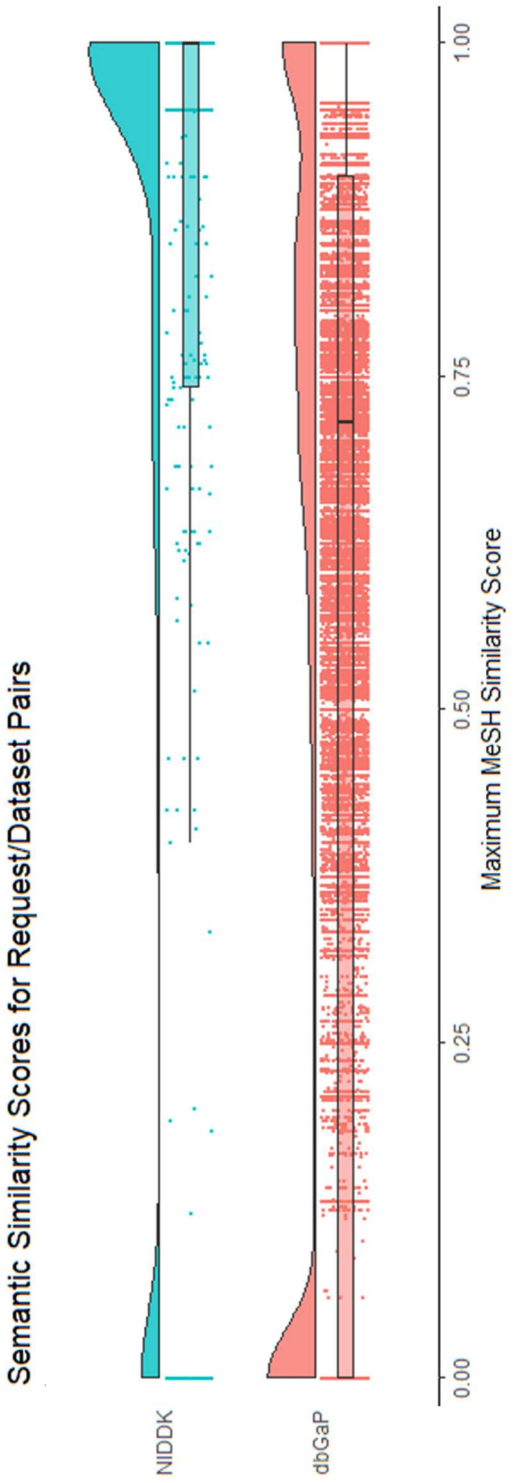

Figure 4-1. Distribution of maximum semantic similarity scores for request/dataset pairs.

A Welch unpaired two-sample t-test shows that the means of the maximum semantic similarity scores for dbGaP and NIDDK request/dataset pairs are significantly different ($t = -14.22$, $df = 546$, $p < 0.001$). These differences suggest that requestors are using dbGaP for topics that vary more from the original data topic than NIDDK requestors. As hypothesized, NIDDK scores tended to be higher (more similar), while dbGaP scores tended to be lower (less similar). Over half of the NIDDK datasets had a score of 1, indicating that requestors intended to reuse the datasets for the same topic of research for which it had originally been collected. On the other hand, nearly a third of dbGaP datasets had a score of 0, suggesting that these datasets were being used in entirely novel contexts compared to the topic for which the data were originally collected.

4.1.3 Summary of Findings

These findings demonstrate that dbGaP and NIDDK datasets are being reused in very different ways from each other. dbGaP datasets were most often used in combination with other datasets to conduct meta-analyses, and they were more likely to be used for a topic that diverged from the original reason the data were collected. On the other hand, just over half of the NIDDK datasets were requested for use in an original research study, using a single dataset on its own. NIDDK datasets were also reused in contexts that were generally more similar to the reason for which the data had originally been collected.

The differences in reuse observed here are likely reflective of the very different types of data in the two repositories. dbGaP houses genetic sequence data; because of statistical issues associated with analyzing this type of data, very large sample sizes are required to achieve adequate statistical power and arrive at meaningful results (Hong & Park, 2012). A number of the dbGaP datasets contain genetic sequences of normal, healthy humans, which can serve as a useful comparison group for a researcher's own set of sequences on a particular disease, since identifying where variations occur in the disease group but not in the healthy comparison group can elucidate genetic regions of interest. In general, genetic sequence data provides more flexibility in its range of research applications than the type of clinical data collected in NIDDK. These clinical datasets tend to be more focused on a specific disease or condition and therefore have less broad applicability. Further, while genetic sequence data is largely standardized and therefore generally interoperable regardless of who collected it, the same is not true for clinical data, which is often recorded based on the specific practices of individual research teams, and therefore more difficult to analyze in combination with other datasets.

4.2 Research Question 2: What are the demographics of researchers who reuse existing datasets?

The repositories included in this study represent a valuable resource for the research community at large, regardless of a researcher's country of origin or career status. A young assistant professor at a small university in South America is just as

eligible to request data as an acclaimed full professor at an Ivy League university. However, just because both of these hypothetical researchers are *able* to request data does not necessarily mean that they *do*. Here, I aim to understand the demographics of researchers who request data by exploring the geographic distribution of requests and the career status of requestors.

4.2.1 Research Question 2.1: Where are requestors located in the world?

Although the three repositories considered here are funded by and administered through various parts of the National Institutes of Health, a United States government research institution, researchers from around the world are permitted to request use of the datasets. While requests can and do come from around the world, I hypothesize that most requests will arise from geographic regions with a large research presence, such as North America, Europe, and Asia, as well as highly-populated states within the US.

Research activities are not distributed evenly among countries around the world, nor among states in the United States. For example, a country such as the United States that is large and has many well-established research institutions is likely to have more dataset requests than a country such as Liechtenstein, which is much smaller and has fewer universities, simply because there are more researchers in the United States to request datasets. Therefore, I calculated relative difference in composition between requests by repository and a proxy measure for presence of research institutions. Relative difference in composition (RDC) is used to quantify

over- and underrepresentation of specific groups in a measure of interest compared to their representation in the population overall (Ford, 2014).

To calculate RDC, first the difference in composition between the measure of interest (requests) and the comparison measure (the proxy measure for research presence) is calculated. For example, suppose that requests from a country constitute 15% of the overall requests to a repository, and that country has 10% of the research institutions in the world. The difference in composition is 5%. Then, the RDC is calculated by dividing the difference in composition by the composition of the research proxy, that is, 5%/10%, and multiplying by 100, yielding an RDC of 200% - that is, that particular country's requests are 200% of what would be expected given its number of research institutions. In this analysis, I use counts of individual requests rather than counts of datasets requested to represent how many studies the repository is supporting. For example, if one researcher from a country is requesting 250 datasets to conduct a single meta-analysis, it is counted as one request, not 250. If each dataset requested was counted individually, a single meta-analysis could significantly sway a country's results, overrepresenting the amount of research supported by the shared data.

A single list of all research institutions of all types globally would be nearly impossible to obtain, so I use number of universities in the country as a proxy for number of research institutions. Although there are various types of non-academic research institutions employing researchers that might request datasets, the number of universities provides a reasonable basis for quantifying the relative research presence

of a given country. The Cybermetrics Lab in the Consejo Superior de Investigaciones Científicas (CSIC), a public research institution in Spain, maintains a list of universities and rankings for 209 countries around the world, including 28,077 universities as of January 2019 (Consejo Superior de Investigaciones Científicas, 2019). I used this list to calculate the percent of all universities in the world located in each country. For example, India has 3,944 universities, the most of any country in the world, accounting for 14% of all global universities. By comparison, a country such as Malawi that has only 12 universities accounts for 0.04% of the world's universities. Considering the difference between the percent of all repository requests coming from a country and the percent of all universities in the world that are in that country provides a basis for determining whether countries are requesting datasets at a rate that is proportional to its representation among global universities.

Figure 4-2, Figure 4-4, and Figure 4-6 show relative difference in composition by each repository internationally. Darker shades of blue indicate more significant underrepresentation of requests relative to number of universities, while darker shades of red indicate more significant overrepresentation. Countries in gray have no universities represented in the CISC list, nor requests to the data repository. Each figure has a different legend based on the maximum difference in relative composition for each repository. Figure 4-3, Figure 4-5, and Figure 4-7 compare counts of universities per country to counts of requests coming from that country for each repository, demonstrating that there is neither a linear nor quadratic relationship between these two variables.

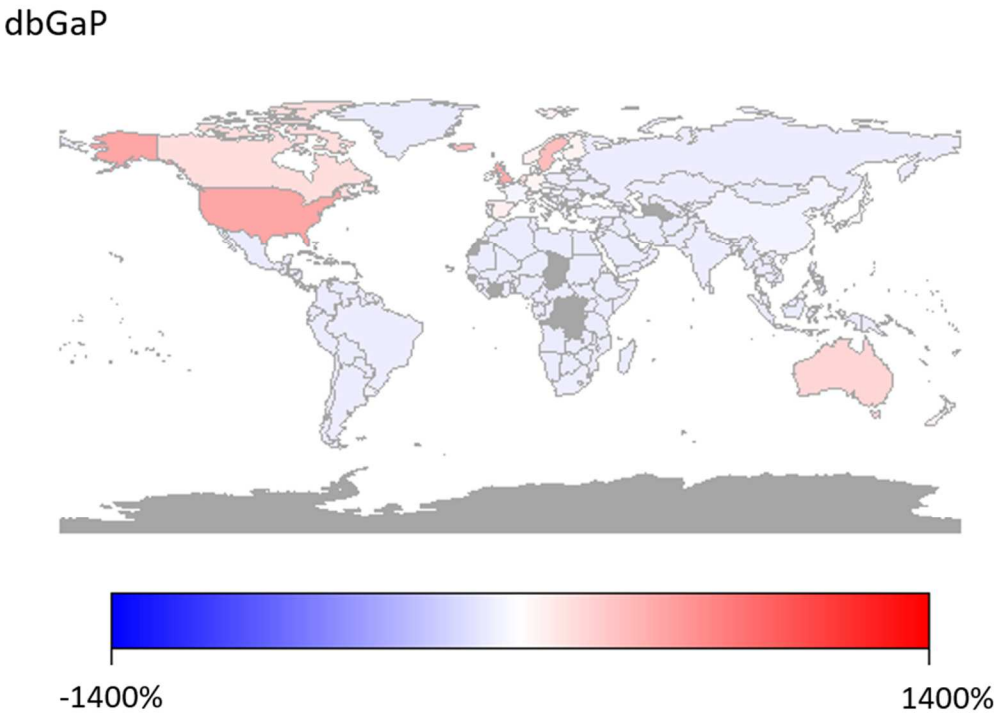

Figure 4-2. Relative difference in composition of requests for dbGaP datasets and universities in countries in the world.

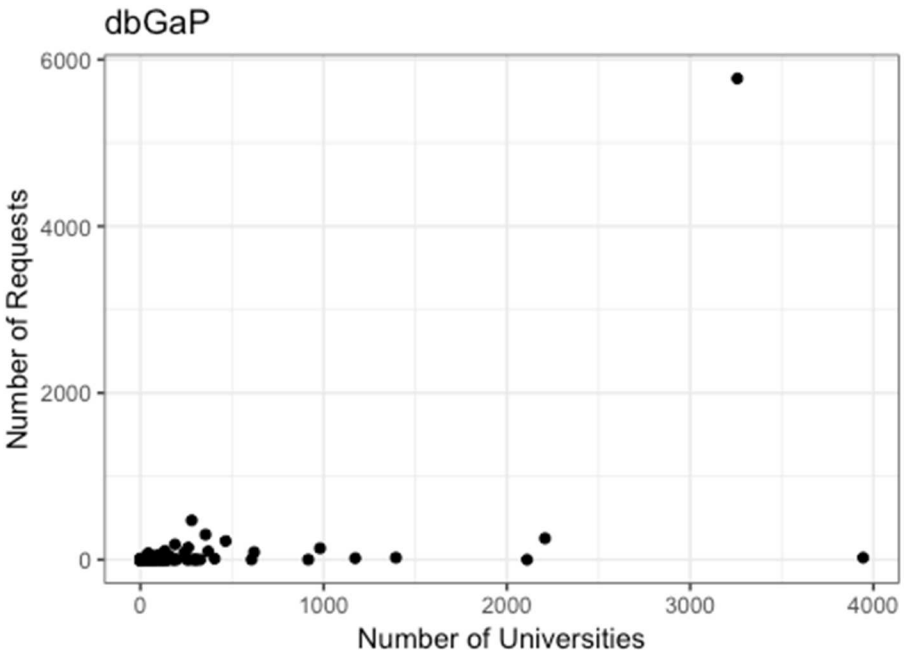

Figure 4-3. Counts of universities compared to counts of requests to dbGaP.

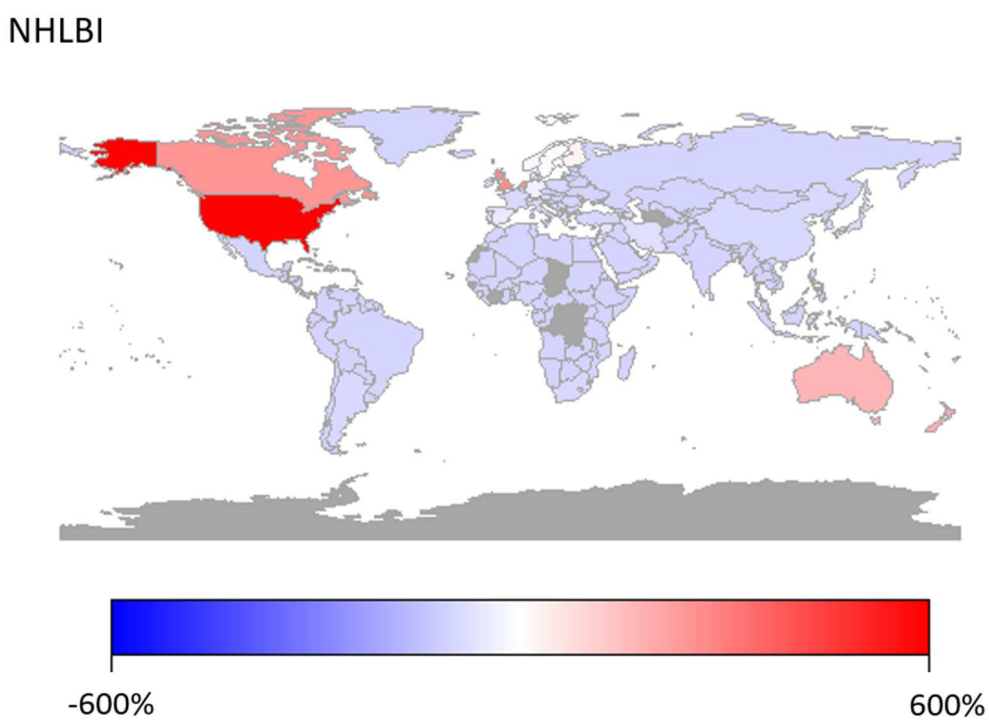

Figure 4-4. Relative difference in composition of requests for NHLBI datasets and universities in countries in the world.

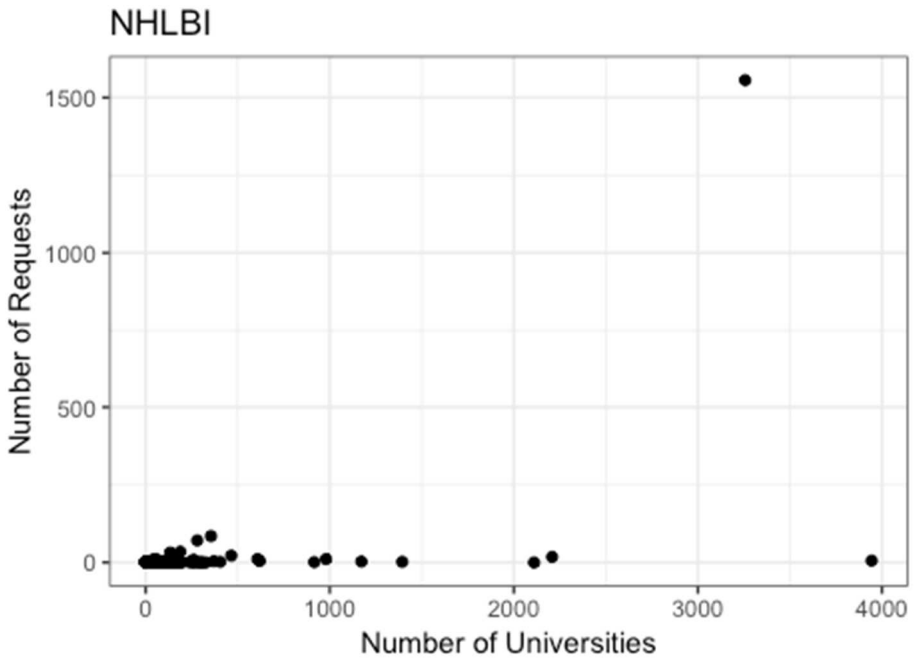

Figure 4-5. Counts of universities compared to counts of requests to NHLBI.

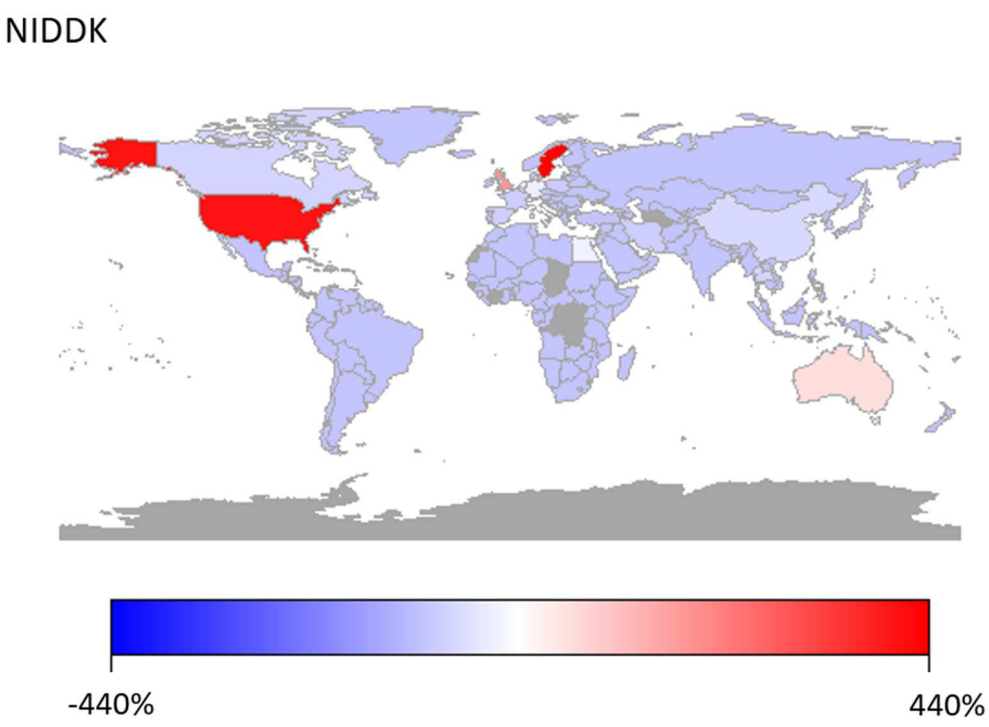

Figure 4-6. Relative difference in composition of requests for NIDDK datasets and universities in countries in the world.

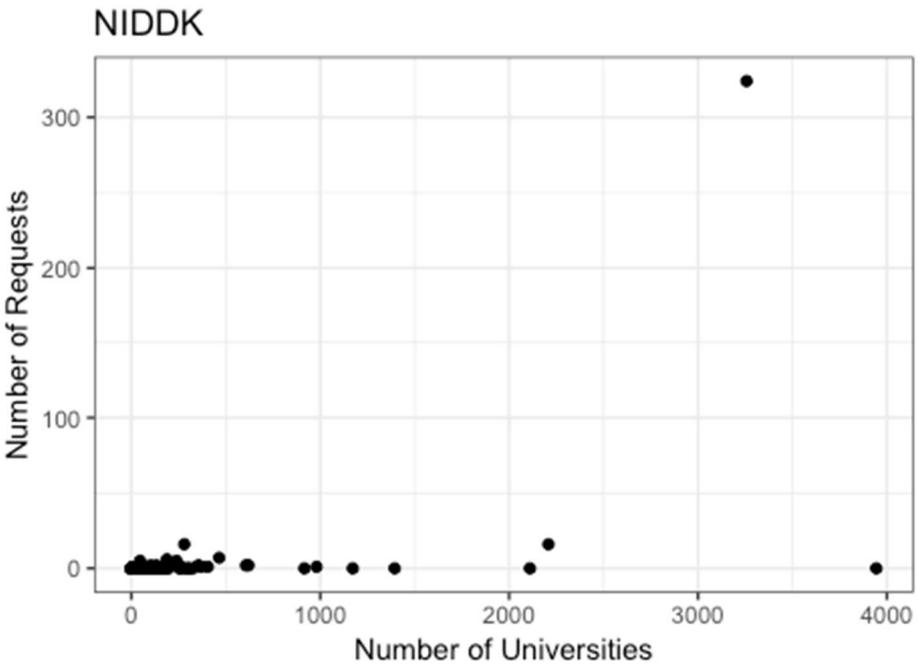

Figure 4-7. Counts of universities compared to counts of requests to NIDDK.

As these three maps demonstrate, requests for datasets are unevenly distributed, with a few countries highly overrepresented. In fact, most countries that had at least one university had never made any requests to the repositories; 79% of countries with universities had no requests to NHLBI, 81% had no requests to dbGaP, and 90% had made no requests to NIDDK. Given that these three repositories are within the United States, it is perhaps unsurprising that United States-based institutions are highly overrepresented among requests from all three repositories. Datasets also appear to be more highly requested in English-speaking countries; Canada, the United Kingdom, and Australia are all over-represented for some or all three of the repositories. This finding could be due to the documentation and web pages of the repositories being written in English; non-English speakers might have

difficulty finding and using datasets that do not include documentation in their native language, especially given that requesting the datasets requires writing a detailed description of the proposed reuse in English.

Table 4-5 shows the number of universities, requests per repository, and relative difference in composition for the ten highest scoring countries for each repository, except NIDDK, which only had six countries that were overrepresented (several countries are in the top ten for more than one repository). RDC values of less than 0 are highlighted in light gray. As Table 4-5 demonstrates, countries were not universally under- or over-represented among requests to the various repositories; in fact, relative difference in composition varied significantly among the repositories. For example, Luxembourg, which had the highest relative difference in composition for dbGaP requests, (1,397% over-represented), did not have one single request to either of the other two repositories and therefore was 100% underrepresented.

Table 4-5. Countries with number of universities and number of requests (N) and relative difference in composition (RDC) for each repository.

Country	University Count	dbGaP		NIDDK		NHLBI	
		N	RDC	N	RDC	N	RDC
Australia	188	183	221%	6	55%	35	170%
Canada	355	301	179%	2	-72%	85	246%
Cyprus	26	1	-89%	1	84%	0	-100%
Finland	46	23	65%	0	-100%	4	28%
Germany	465	223	58%	2	-26%	22	-32%
Iceland	9	12	337%	0	-100%	0	-100%
Israel	42	77	501%	0	-100%	10	248%
Italy	239	86	19%	5	2%	1	-94%
Luxembourg	3	14	1,397%	0	-100%	0	-100%
Netherlands	133	106	162%	2	-26%	32	248%
New Zealand	56	27	60%	0	-100%	11	186%

Country	University Count	dbGaP		NIDDK		NHLBI	
		N	RDC	N	RDC	N	RDC
Qatar	9	0	-100%	0	-100%	1	56%
Singapore	45	44	224%	0	-100%	3	-6%
Sweden	46	63	352%	5	431%	3	-8%
Switzerland	102	59	90%	2	-4%	4	-42%
United Kingdom	280	471	484%	16	179%	71	267%
United States	3,257	5,773	484%	338	406%	1,556	592%

Among the most highly overrepresented countries, the large number of requests cannot be explained by coming from one highly prolific requestor or institution. For example, all of the dbGaP requests from Luxembourg do come from just one of its three national universities, but the 14 requests come from nine different requestors. The 77 requests to dbGaP from Israel, the next most overrepresented country, come from 15 different institutions. However, some of the countries that are overrepresented in fact have a low number of requests and only appear overrepresented because they also have very few universities. For example, Qatar is the eighth most highly represented country among NHLBI requests despite having only one request. In fact, 27 countries have more requests than Qatar, but 20 of them have a lower RDC because of the much larger number of universities they have than Qatar's nine.

Just as research institutions are not evenly distributed around the world, they also are not within the United States among states. I conducted the RDC analysis for states as well, using NIH funding amounts in Fiscal Year 2018 (National Institutes of Health Research Portfolio Online Reporting Tools, 2018) as a proxy for research

presence. I calculated the relative difference between the percent of total requests by repository made in a state and the percent of all NIH funding awarded within the United States that was awarded to that state. NIH research funding is probably a more accurate proxy for biomedical research presence than the university count that was feasible to use for the world analysis, since NIH awards funding to a variety of types of research institutions, not just universities, and focuses specifically on the type of biomedical research that is relevant here.

Figure 4-8, Figure 4-9, and Figure 4-10 show RDC by repository within the United States. Red indicates states that are requesting a larger share of datasets compared to the research funding they receive, while blue indicates states that are requesting a smaller share. The darker the color, the more highly the state is over- or underrepresented, while states in white request datasets at a rate about equivalent to their research presence.

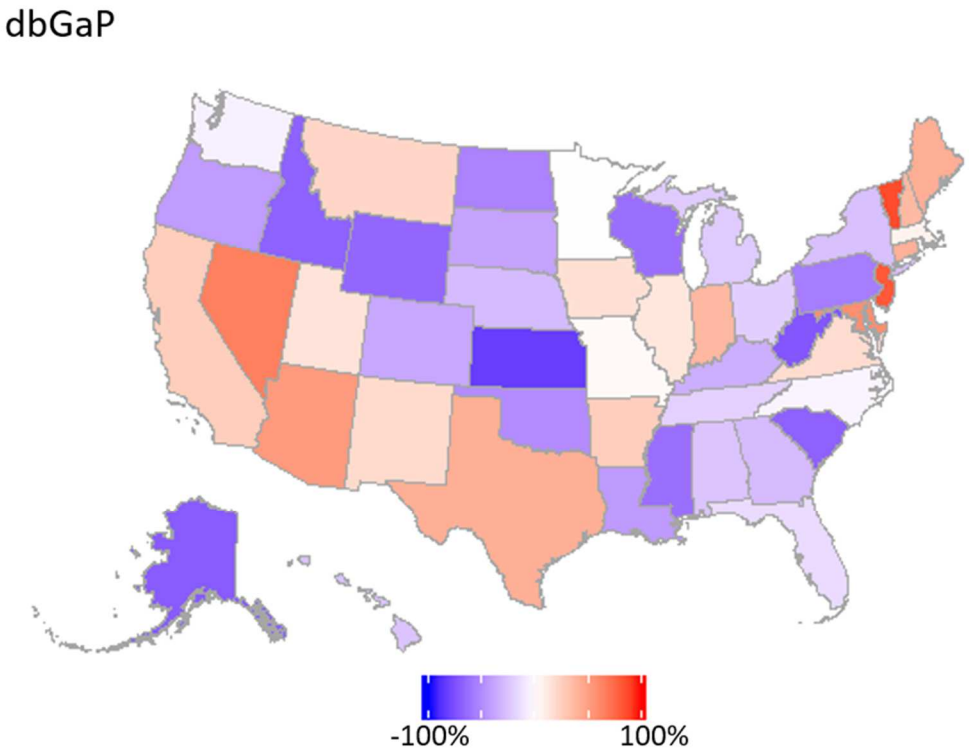

Figure 4-8. Relative difference in composition of requests for dbGaP datasets and NIH funding in FY18 by state within the US.

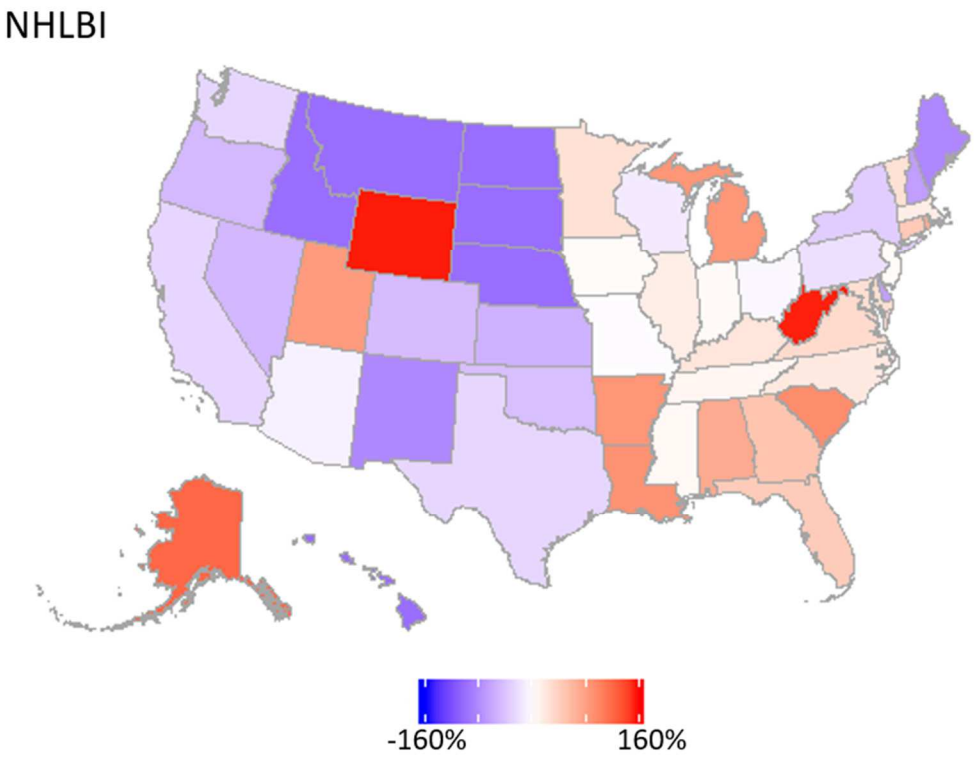

Figure 4-9. Relative difference in composition of requests for NHLBI datasets and NIH funding in FY18 by state within the US.

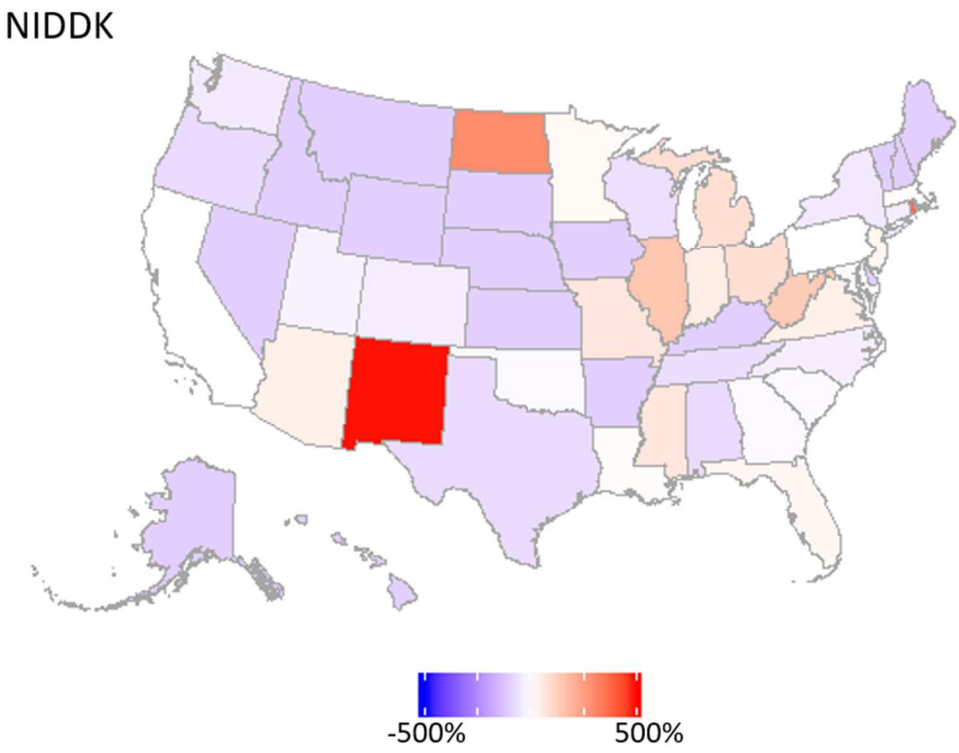

Figure 4-10. Relative difference in composition of requests for NIDDK datasets and NIH funding in FY18 by state within the US.

The state RDC analysis shows more variation in geographic distributions than the global RDC analysis. The states that are the most highly over-represented among the various repositories are not necessarily the ones that might be expected: New Mexico, Wyoming, and Alaska all appear as outliers. On the other hand, other states with a strong research reputation also are over-represented, such as Massachusetts and California. Unlike the global analysis, more states appear in white (or a shade close to it), indicating that they are requesting datasets at a level that is proportional to the amount of NIH funding they receive. This finding could suggest that requests for data are more evenly distributed among research institutions within the United States

than they are within universities across the world. Where disparities do exist within the states, they also generally tended to be less significant than those among countries. Compared to RDCs of nearly 1,500% for the most highly overrepresented countries, the most extreme RDCs for NHLBI is about 500% and dbGaP's is only 85%. However, NIDDK requests are skewed at a level closer to that seen at the global level, largely due to the very high overrepresentation of requests from New Mexico and Washington, DC.

As with the global RDC analysis, some states appear highly overrepresented not because they have a very high number of requests, but because they receive very little NIH funding. For example, only two requests for NHLBI data came from Wyoming, but they also receive the least NIH funding of any state – only 0.05% of all NIH funding. However, some highly funded states also request very little data. For example, Texas, the seventh-highest funded state, had only made four requests to NIDDK.

4.2.2 Research Question 2.2: Are there patterns in career stage of requestors?

Although requestors to the three repositories must demonstrate that they are legitimate researchers (for example, dbGaP requestors must be registered in NIH's Electronic Research Administration system, while NHLBI and NIDDK have a process for requestors to apply for an account, which includes indicating their research affiliation and status), researchers from a range of career stages are free to request datasets. That range includes students to full professors, as well as career

stages from areas outside of academia, such as senior scientists, CEOs and other executives, and managers. Some requestors may be at a career stage at which they might benefit more substantially from the opportunity to use existing data – for example, students and early career researchers are less likely to have access to the significant funding, laboratory resources, and staff that it would take to generate their own data. Despite the potentially greater benefit to early career researchers, I hypothesized that a broad range of career stages, from student to full professor (or their equivalents in non-academic contexts) would be represented.

For this analysis, I used the NIDDK requests and a random sample of 1,500 of the total 9,444 dbGaP requests (15.9%), which provides a +/-1.1 confidence interval at a 95% confidence level based on estimation of proportion. Of the 416 NIDDK requests, 144 of them (35%) did not include a requestor name and were therefore excluded from this analysis, leaving 272 requests. NHLBI did not provide me individual researcher level request data for privacy reasons, so those requests could not be included in this analysis. I determined the career status of the researcher at the time they made the request by searching the internet for documentation of their career history, such as institutional web pages, CVs, biosketches, and LinkedIn pages. Titles from non-American institutions were converted to their American equivalent; for example, the rank of "senior lecturer" in the United Kingdom is the equivalent of an associate professor in the US (Wikipedia, 2018). For requestors for whom I could not definitively determine the career status at the time of request, I recorded "unknown."

The 1,500 dbGaP requests came from 1,118 unique requestors and requested access to 18,117 total datasets (since a request could ask for multiple datasets). Each unique request asked for between one and 529 datasets, with a mean of 12.1 datasets per request. The 272 NIDDK requests came from 252 unique requestors, requesting a total of 394, with each request asking for between one and ten datasets (mean 1.4). While many requests asked for more than one dataset, this analysis counts individual requests rather than requests by dataset to provide a clear understanding of how much research is being supported at each career stage. For example, if an associate professor requests 30 datasets for a meta-analysis, that request supports one research project; counting each dataset separately would inflate counts of how much research is being supported at a given career stage. Table 4-6 provides the distribution of requests for dbGaP and NIDDK, by career status, and with statuses grouped by career stages that approximately reflect where the career status falls in a broader career trajectory.

Table 4-6. Proportions of datasets requested by career status of requestor for dbGaP and NIDDK.

Career Stage	Title	Percent of dbGaP requests	Percent of NIDDK requests
Pre-professional	Student	0.7%	1.8%
	Fellow	0.7%	3.1%
	Total	***1.4%***	***4.9%***
Early career	Assistant Professor	19.1%	27.6%
	Resident Physician	0%	1.1%
	Lecturer	0.07%	0.4%
	Instructor	0.07%	0%
	Total	***19.2%***	***29.1%***
Mid-Career	Associate Professor	15.4%	13%
	Scientist	5.7%	3.9%

Career Stage	Title	Percent of dbGaP requests	Percent of NIDDK requests
	Attending Physician	0%	0.2%
	Manager	0.7%	0.4%
	Total	***21.8%***	***17.5%***
Established	Professor	26.8%	24%
	Director	8.5%	5.5%
	Executive	3%	5.1%
	Senior Scientist	10.3%	6.7%
	Total	***48.6%***	***41.3%***
Unknown		**9%**	**5.9%**

Patterns of requests appeared generally similar across dbGaP and NIDDK; however, a chi-squared test of independence revealed that the two distributions were in fact significantly different ($\chi 2 = 81$, df = 12, $p < 0.001$). This statistic was most influenced by the numbers of resident and attending physicians requesting datasets; expected counts would be 1 physician (0.9 expected resident and 0.1 expected attending) for NIDDK and 6 for dbGaP (5.1 expected resident and 0.9 expected attending), but all 7 requests from physicians went to NIDDK. This finding could be explained by the fact that NIDDK contains clinical data of the type that would be familiar to physicians, whereas physicians generally do not have training in dealing with genomic information and would be therefore be less likely to use the genomic data found in dbGaP (Demmer & Waggoner, 2014; Manolio & Murray, 2014; Murray, 2014).

Despite the fact that the distribution of requestors between the two repositories differed statistically, the requests did at least follow a broadly similar pattern, with nearly half of requests to both repositories coming from full professors

and other researchers in more established positions. Assistant professors also represented a sizeable proportion of requestors, accounting for about a quarter of the datasets requested from both dbGaP and NIDDK. Almost none of the requests came from pre-professionals such as students and fellows. However, a limitation that should be noted for this analysis is that the person who requested the data might not be the person who actually ended up *using* the data. For example, a full professor might request data on behalf of a graduate student.

As with universities' uneven distribution around the world, researchers are not necessarily evenly distributed among career ranks. For example, faculty might be more concentrated in lower ranks, and therefore it would be expected that they would make more requests, since there are more individuals to be making requests. Therefore, in addition to considering proportions overall, I also calculated the relative difference in composition (RDC), as described in Section 4.1. Obtaining counts of non-academic ranks such as CEO or scientist was infeasible, but I calculated RDC for the academic-related ranks based on 2016 data from the National Center for Education Statistics, which reports counts of full-time faculty in US degree-granting postsecondary institutions (National Center for Education Statistics, 2017). I compared the proportion of each rank within all of US faculty to its proportion of academic requests for dbGaP and NIDDK. Note that this analysis only considers requests that came from academic requestors; for example, the 46.3% reported for professors requesting dbGaP datasets refers not to the proportion of datasets this group requested compared to *all* requests, but to the proportion requested compared

to requests coming from the five academic ranks listed in the results, reported in Table 4-7.

Table 4-7. Relative difference in composition (RDC) between faculty at five academic ranks in US institutions and their requests to dbGaP and NIDDK.

Faculty status	Percent of US faculty	Academic dbGaP requests		Academic NIDDK requests	
		%	RDC	%	RDC
Professor	22.4%	44%	96%	40%	78%
Associate professor	19.3%	25%	29%	20%	4%
Assistant professor	21.6%	31%	43%	42%	94%
Instructor	12.4%	0.1%	-99%	0%	-100%
Lecturer	5.2%	0.1%	-98%	0.6%	-88%
Other	19.1%	NA	NA	NA	NA

A chi-squared test of independence revealed that request counts from staff at different faculty ranks differed significantly from their representation in American universities for both dbGaP and NIDDK ($\chi2 = 641$, df = 5, $p < 0.001$ and $\chi2 = 108$, df = 5, $p < 0.001$, respectively). Their distributions are also significantly different from each other ($\chi2 = 14$, df = 4, $p = 0.01$). As Table 4-7 demonstrates, professors are overrepresented in their requests to both repositories, although to a lesser degree among requests to NIDDK. Instructors and lecturers are almost 100% underrepresented, a finding that seems reasonable given that many faculty members at this level have teaching and service responsibilities that may limit their engagement in research, and therefore request less data for that purpose.

A surprising finding is that the representation of assistant professors and associate professors varies between dbGaP and NIDDK. Associate professors are 30% overrepresented among dbGaP requests but only 4% represented among NIDDK

requests. Assistant professors, on the other hand, are 94% overrepresented among NIDDK requests, but less than half as much overrepresented among dbGaP requests (43%). A possible explanation for this finding could be that researchers at different ranks are more likely to engage in the types of research that tend to be supported by each repository, that is, that associate professors are requesting more data from dbGaP because they are doing more meta-analyses and assistant professors are requesting more data from NIDDK because they are doing more original research studies. Further research into how requestors are using datasets could help elucidate some of the differences in request rates.

4.2.3 Summary of Findings

Although datasets from the three repositories considered here are theoretically available to any qualified researcher, requests for datasets are unequally distributed around the world and among researchers at different career stages. English-speaking regions, particularly the United States, were overrepresented in requests compared to their number of research institutions. Established researchers who were at higher career ranks were also overrepresented, particularly among academic staff. These findings suggest that, in many cases, datasets are going to the researchers most able to collect their own data if need be: established researchers in wealthy countries who likely have access to resources that earlier career researchers and those in poorer countries do not.

4.3 Conclusions and Summary of Findings

The results reported here have helped elucidate the who, where, and why of data reuse. From these findings, a general picture of biomedical data reuse begins to appear. Researchers are making use of data in a wide range of contexts, from using one dataset in a context very similar to its original purpose, to requesting hundreds of datasets from a range of unrelated topics to conduct large-scale meta-analyses. The range of types and contexts of reuse seen here demonstrates that data reuse is complex, not a single, easily explained phenomenon, although some of the differences in reuse can be explained by the repository and the type of data it holds. Researchers from around the world are taking advantage of the opportunity to reuse existing datasets rather than gathering their own, though requests tend to be concentrated in English-speaking countries, particularly the United States. Requests come from researchers at all different career stages, from students just beginning their career to full professors who are well established in their discipline, though later career researchers are somewhat overrepresented. In Chapter 5, I will build on this emerging picture of biomedical data reuse by considering patterns of use requests in relation to dataset topic and time since dataset release.

Chapter 5: Findings About Datasets

This chapter presents the findings of the two research questions that focus on questions about the datasets themselves, or the when and what of biomedical data reuse – when in a dataset's life cycle is it most requested, and what topics are the most requested? Specifically, the research questions and hypotheses considered here are:

Research Question 3: Are there temporal patterns to dataset requests?

> **Hypothesis 3**: Patterns of requests relative to the original dataset release date will demonstrate a cumulative advantage process, similar to other scientific communication processes such as article citation.

Research Question 4: Are there dataset topics that are more highly requested?

5.2 Research Question 3: Are there temporal patterns to dataset requests?

Many processes in the study of science, including citations to articles, follow the model of a cumulative advantage process: the rich get richer, and success breeds success. In other words, an article that has already been cited many times is more likely to go on to receive more citations than an article that has only been cited a few times. This process makes sense for a variety of reasons – an article cited many times could be cited more because it is of higher quality than a less-cited article, and a highly cited article likely ends up having more visibility than a less-cited article, since it appears in the bibliography of more citing articles. I hypothesize that temporal

patterns in requests for datasets over time can, like article citations, be explained by a cumulative advantage model.

For this analysis, I used dbGaP and NHLBI datasets only. The NIDDK repository only had a specific release year for datasets released in 2014 or later, which is just 30% of its datasets. As a result, only 91 of the total 516 requests, just 18%, could be matched with a dataset with a known release date, and most datasets had only one or two requests per year. With so few datasets and only four years' worth of requests to consider, the NIDDK data was inadequate for this analysis.

Request data began in 2007 for dbGaP and 2000 for NHLBI. For both the dbGaP and NHLBI analysis, 2018 requests were excluded since the list of dataset requests was collected in mid-2018 and therefore did not represent a full year worth of requests. Thus, the dbGaP analysis included requests made between 2007 and 2017, and the NHLBI analysis, requests made between 2000 and 2017.

For each repository, I considered how many total requests each dataset had received across the years included in this analysis (that is, excluding 2018 requests). Based on these total requests, I determined rankings for the least to the most requested datasets by calculating how many requests a dataset would need to fall in each decile (or set of 10 percentile points) between the 10th and 90th percentile and determined the decile for each dataset. For example, a dbGaP dataset with a total of 6 requests would be in the 20th percentile, while a highly requested dataset that had received 200 requests would be in the 90th percentile. For each individual request, I determined the age of the dataset at the time of the request by subtracting the year the

request was made from the year the dataset was first shared. Finally, I calculated the mean number of requests by decile for each year since dataset release (for example, of the 365 datasets that fall into the 50th percentile for dbGaP, the mean number of requests they received in the first year of being available was 5.85).

Using the dataset's age at the time of request (e.g. the request was made 2 years after the dataset was released) rather than the calendar year of the request (e.g. the request was made in 2015) makes it possible to compare datasets of different ages. If the cumulative advantage effect holds true, a dataset released in 2009, for example, would be more likely to have a higher number of requests in 2015 than a dataset released in 2014, since it is six years old and has had more time to accumulate advantage than a one-year-old dataset. However, the number of requests the 2009 dataset received in 2010, when it was one year old, can be reasonably compared to the number of requests the 2014 dataset received in 2015, when it was also one year old.

Of course, it is possible that the year a dataset was released might affect the number of requests it receives even when comparing like to like by using dataset age. For example, data science and other computational methods have become increasingly popular in recent years, so perhaps a dataset released in 2015 would be more requested in its first year than a dataset released in 2009 would in its first year, simply because more people are making requests overall. However, this analysis also controls for age, as will be further discussed, by measuring correlation between year of release and total requests. A weak correlation between year of release and number

of requests received would suggest that year of release has little impact on a dataset's number of requests.

5.1.1 dbGaP Results

The dbGaP analysis includes 982 datasets with a total of 100,115 requests between 2007 and 2017; 68 datasets for which a year of release could not be determined from the dbGaP website were excluded. Figure 5-1 shows the number of requests datasets in each decile received in each year since their release (not cumulative requests). The count of age at request on the x-axis begins with 0, which indicates requests made within the first year of its release, with 1 indicating requests when the data is one year old, and so on. Table 5-1 shows the range and distribution of dbGaP datasets within deciles.

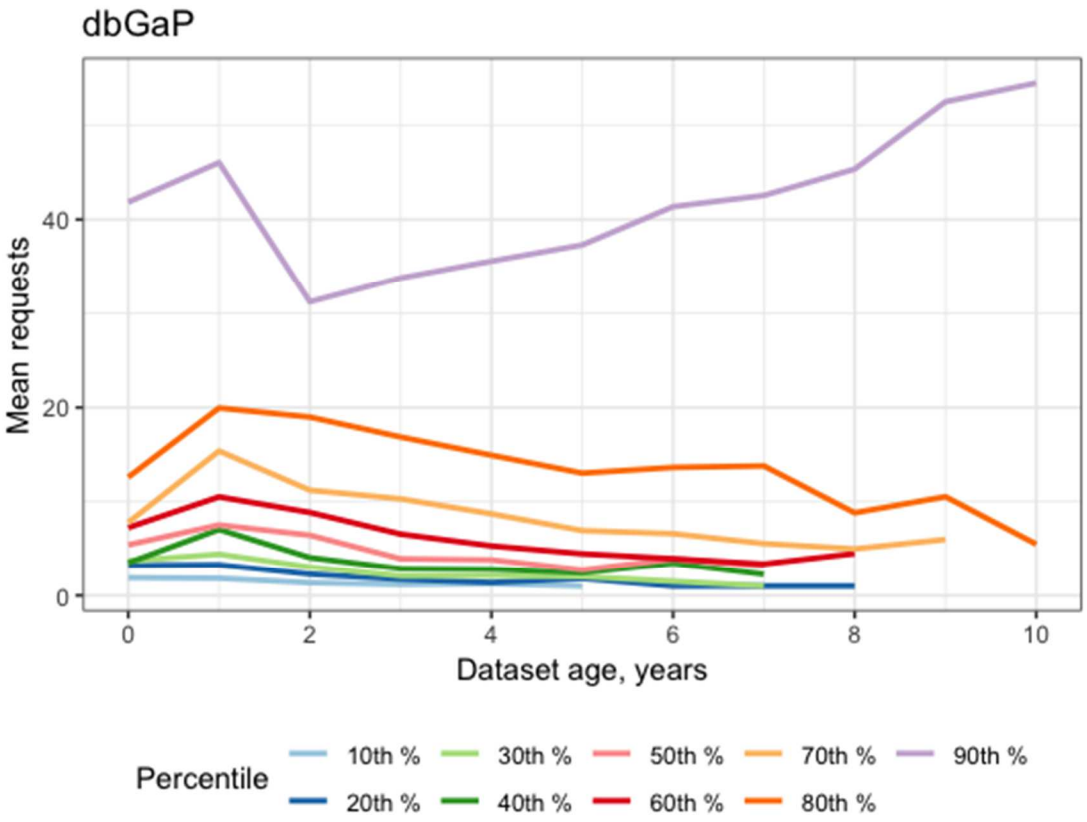

Figure 5-1. Mean requests by year for dbGaP datasets in each decile, by age of the dataset at time of request.

Table 5-1. Distribution of dbGaP datasets by request deciles for requests made between 2007 and 2017.

Decile	Request count range	Number of datasets in percentile	Mean age of datasets, years
10th percentile	4 or fewer	164	1.82 (range 0 –7)
20th percentile	5 – 8	263	2.89 (range 0 – 8)
30th percentile	9 – 12	254	3.57 (range 0 – 7)
40th percentile	13 – 18	324	3.85 (range 0 – 7)
50th percentile	19 – 27	373	4.42 (range 0 – 7)
60th percentile	28 – 42	406	4.66 (range 0 – 8)
70th percentile	43 – 66	465	5.16 (range 1 – 9)
80th percentile	67 – 174	635	6.75 (range 1 – 10)
90th percentile	175 – 2754	1807	9.03 (range 5 – 10)

As Table 5-1 demonstrates, datasets in the lower percentiles of requests are on average younger, which is logical considering that a ten-year old dataset has had twice as long to accumulate requests as a five-year old dataset, and would therefore be in a higher decile. It does appear that length of data availability does at least partly explain the amount of requests a dataset has received; no datasets that had less than five years to accrue citations (i.e. no datasets released after 2012) made it into the 90th percentile, and none of the oldest datasets (released between 2007 and 2009) fell below the 80th percentile. However, a dataset's age cannot fully account for the number of requests it has received, given that at least some datasets that had eight years to accrue requests were only in the 20th percentile, compared to other datasets that had only one year to accrue requests and made it into the 60th percentile.

Further, as discussed above, this analysis controls for age by comparing requests over time based on dataset age rather than calendar year. Datasets in the 90th percentile were already more highly requested in the first year after being released, receiving on average 42 requests in the first year – more than three times as many as datasets in the 80th percentile received in their first year (mean = 13) and more than twenty times as many as datasets in the bottom 10th percentile received in their first year (mean = 2). The number of requests in the first year varies significantly even among the lower deciles; on average, datasets in the 20th percentile received over 70% more requests in their first year than those in the 10th percentile.

Because this method is still somewhat affected by the age of the dataset, I also calculated percentile ranges for each year of release (i.e. calculated percentiles for all

requests in year one of a dataset's life, year two, and so on) and conducted the same analysis using the mean percentile across all the years of its availability, rather than its overall percentile. For example, a dataset in the 90th percentile of first-year requests, 80th percentile of second-year requests, and 70th percentile of third-year requests would have a mean of the 80th percentile. Using the mean instead of the overall percentile ranking helps compare newer datasets more fairly with older datasets. Whereas an older dataset would be more likely to be in a higher percentile overall because it had more time to accrue requests, using the mean percentile makes it possible to compare datasets at various points in their life. For example, suppose a dataset has been available for two years, and receives 25 requests in its first year and 40 requests in its second year, putting it in the 90th percentile for both first- and second-year requests, for a mean of the 90th percentile. In its two years, it has accrued a total of 65 requests, but this is only enough to put it in the 70th percentile overall, since it is also competing against datasets that have had five times as long to accrue total requests. Using the mean percentile instead of the overall percentile more accurately reflects its high performance over the course of its life, comparing it to datasets of the same age at each year of its life.

Most datasets had at least some variability in their performance across all years, with few achieving the same percentile at every year since their release. Because of this variability, and since the mean is a measure of central tendency, few datasets had a mean percentile at the highest end – only two dataset older than two years old had a mean of 90th percentile. Therefore, rather than use deciles, I grouped

more datasets together and use quartiles (i.e. 0- 25th percentile, 26-50th percentile, 51-75th percentile, and 76-100th percentile). Figure 5-2 shows the number of requests for datasets in each mean quartile received in each year since their release (not cumulative requests). The count of age at request on the x-axis begins with 0, which indicates requests made within the first year of its release, with 1 indicating requests when the data is one year old, and so on. Table 5-2 shows the range and distribution of dbGaP datasets within deciles. As Table 5-2 demonstrates, calculating the mean quartile by averaging the percentile for a dataset in each year of its life creates a more even distribution of datasets by age among the various quartiles. However, the two higher quartiles do have older datasets on average than the two lower quartiles.

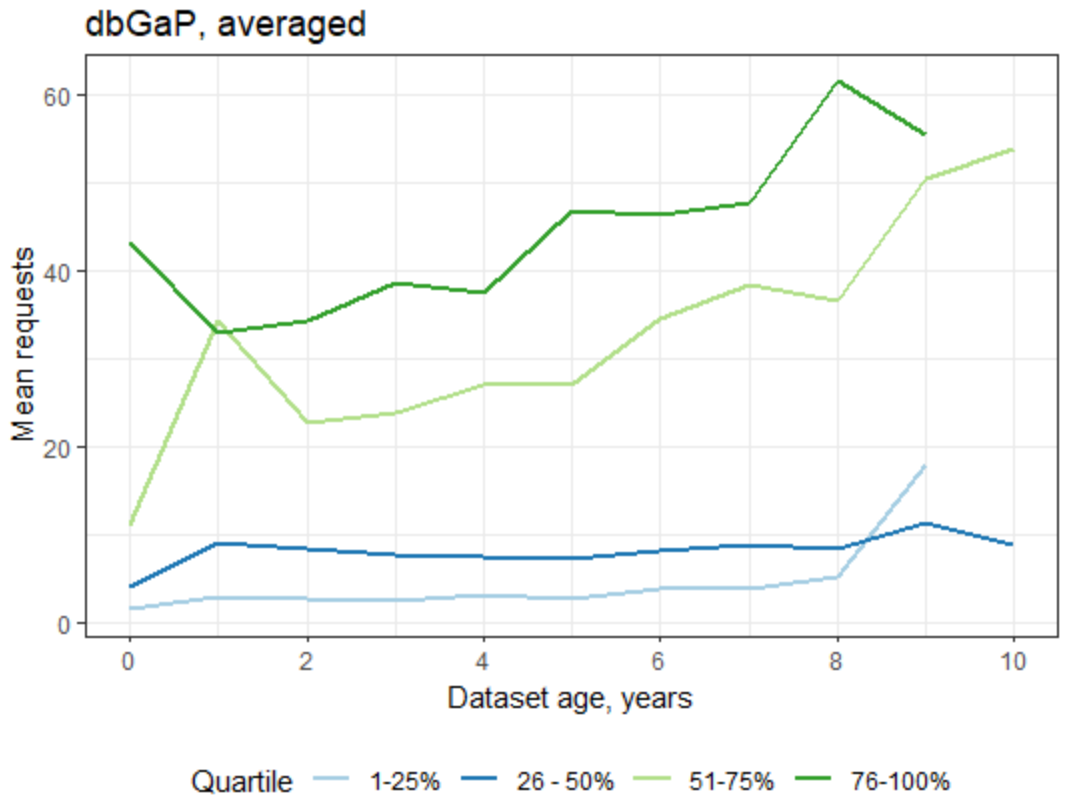

Figure 5-2. Mean requests by year for dbGaP datasets in mean quartile, by age of the dataset at time of request.

Quartile	Total request count range	Number of datasets in percentile	Mean age of datasets, years
1-25th percentile	1 – 114	867	4.7 (range = 0-9)
26-50th percentile	2 – 215	1321	4.9 (range = 0-10)
51-75th percentile	5 – 442	1808	7.7 (range = 0-10)
76-100th percentile	10 – 2754	695	7.8 (range = 0-9)

Table 5-2. Distribution of dbGaP datasets by mean request quartiles for requests made between 2007 and 2017.

While the exact dynamics of requests in the analysis using the mean percentile rather than the overall percentile differ, as Figure 5-2 demonstrates, the general pattern is the same: datasets that start out being highly requested go on to continue

being highly requested as time goes on. Taken together, the mean percentile and overall percentile analyses suggest that dbGaP dataset requests are at least partly affected by a cumulative advantage process. Datasets that are highly requested soon after their release go on to continue to receive more requests later, while datasets that initially receive fewer requests continue to be less requested over time.

As Figure 5-1 demonstrates, datasets across all overall deciles reach a peak of requests in their second year (age = 1) and requests begin to drop off in the third year. This pattern is very similar to citations to articles over time – articles reach a peak of citations at various ages depending on discipline (for example, Clinical Medicine articles peak around 4 years while Biology articles peak around 7 years), but eventually drop off as the articles becomes older (Eom & Fortunato, 2011; Parolo et al., 2015; Wang, 2013). As with article citations, this decline continues over time for datasets in the 10th through 80th percentiles overall, but the same is not true of datasets in the 90th percentile overall. After following the pattern of third year drop-off, requests actually begin to increase again in the fourth year and steadily climb in each subsequent year, eventually reaching and even surpassing the number of requests received in the first year. With only ten years of data to consider here, it is difficult to completely explain the mechanism behind this pattern. Perhaps these highly requested datasets see a bump in requests in subsequent years as early requestors begin to publish articles that cite their reuse of the dataset, thus creating a cycle of increased attention for the already highly requested datasets.

Looking at the data requests as a whole, rather than dividing them by deciles, also demonstrates a strong relationship between the number of requests a dataset receives in the first few years after release and the number it receives over the long term. The number of long-term requests is strongly positively correlated with the number of requests received in the first year and second year (correlation coefficient for both = 0.8), and even more so with requests in the third year (correlation coefficient = 0.9). There is a negative correlation between release year (i.e. calendar year) and total number of requests, indicating that older datasets have more requests, but this correlation is only moderate (correlation coefficient = -0.6).

Fitting a linear regression model to the request data further demonstrates the importance of a dbGaP dataset's early performance in predicting its long-term request rate. Table 5-3 summarizes results from three regression models: first-year requests only; first- and second-year requests and first-, second-, and third-year requests. All three models also include release year to control for the influence of a dataset's age on the number of requests. All three models are statistically significant at $p < 0.001$. To determine whether results were affected by collinearity, I calculated the variance-inflation factor (VIF) for each model and each variable. VIF values of greater than 10 indicate collinearity; all VIFs here are less than 10 (Dormann et al., 2013).

The R-squared value of a regression model is a measure of the amount of variability in the outcome variable (total requests) that is explained by a given model. For example, the one-year model accounts for 73% of the variability in total requests, while adding the second and third year increases the amount of variability the model

accounts for. The regression coefficient (coef) represents the mean change in total requests for every additional increase of one in the predictor variable, while holding other variables constant. For example, in the one-year model, the coefficient of 6.61 for number of requests in the first year means that for every additional request in the first year, a dataset would have, on average, 6.61 additional requests over time. The standard error (SE) is a measure of the average distance between the regression line and the values in the data. The higher the standard error, the less correct the model is on average; variables with SE of greater than 2.5 are not statistically significant at a 95% prediction interval. Finally, the Beta value (β) indicates the relative influence (and the direction of that influence) of a variable on the number of total requests a dataset receives. Values close to 1 (or -1) indicate a high level of influence, while values close to 0 indicate less influence.

Table 5-3. Results of regression analysis showing effects of requests during year one, two, and three of a dbGaP dataset's life on the total number of requests during the 2007 – 2017 period.

	One-Year Model	**Two-Year Model**	**Three-Year Model**
First year requests	coef = 6.61 ($p <$ 0.001) SE = 0.26 ($p <$ 0.001) β = 0.76 ($p <$ 0.001) VIF = 1.1	coef = 4.6 ($p <$ 0.001) SE = 0.28 ($p <$ 0.001) β = 0.53 ($p <$ 0.001) VIF = 1.11	coef = 3.54 ($p <$ 0.001) SE = 0.23 ($p <$ 0.001) β = 0.41 ($p <$ 0.001) VIF = 1.6
Second year requests	NA	coef = 5.44 ($p <$ 0.001) SE = 0.45 ($p <$ 0.001) β = 0.38 ($p <$ 0.001)	coef = -0.56 ($p <$ 0.001) SE = 0.51 ($p <$ 0.001) β = -0.04 ($p <$ 0.001)

	One-Year Model	Two-Year Model	Three-Year Model
		VIF = 2	VIF = 4
Third year requests	NA	NA	coef = 7.57 (p < 0.001) SE = 0.46 (p < 0.001) β = 0.58 (p < 0.001) VIF = 4.9
Year of release	coef = -10.92 (p < 0.001) SE = 2.31 (p < 0.001) β = -0.14 (p < 0.001) VIF = 1.1	coef = -5.42 (p < 0.001) SE = 2.05 (p < 0.001) β = -0.07 (p < 0.001) VIF = 2.04	coef = -5.8 (p < 0.001) SE = 1.63 (p < 0.001) β = -0.08 (p < 0.001) VIF = 2.2
R-squared	0.733 (p < 0.001)	0.7979 (p < 0.001)	0.8733 (p < 0.001)

As Table 5-3 shows, the three-year model accounts for nearly 90% of the variability in long-term requests. Even the model with only first-year requests accounts for 73% of the variability in total requests. Year of release appears to have only a small amount of influence on the total number of requests, with Beta values close to 0 for all three models. These models suggest that the number of requests a dataset receives in the first few years is a good predictor of long-term requests, regardless of when the dataset was released.

Because first, second, and third year requests are included in total requests, I also fit models to determine the effect of first through third year requests on all later requests, in other words, total requests made in the fourth year and beyond. This analysis includes the 615 datasets that were released before 2015 (that is, those that were old enough to have more than three years' worth of requests). All three models

also include release year to control for the influence of a dataset's age on the number of requests. All three models are statistically significant at $p < 0.001$. Table 5-4 summarizes the results of these three models.

Table 5-4. Results of regression analysis showing effects of requests during year one, two, and three of a dbGaP dataset's life on the total number of requests in the fourth year and later during the 2007 – 2017 period.

	One-Year Model	**Two-Year Model**	**Three-Year Model**
First year requests	coef = 2.53 ($p < 0.001$) SE = 0.22 ($p < 0.001$) β = 0.32 ($p < 0.001$) VIF = 1.1	coef = 2.54 ($p < 0.001$) SE = 0.16 ($p < 0.001$) β = 0.32 ($p < 0.001$) VIF = 1.1	coef = 1.16 ($p < 0.001$) SE = 0.18 ($p < 0.001$) β = 0.14 ($p < 0.001$) VIF = 1.78
Second year requests	NA	coef = 4.75 ($p < 0.001$) SE = 0.2 ($p < 0.001$) β = 0.65 ($p < 0.001$) VIF = 2.21	coef = 2.18 ($p < 0.001$) SE = 0.27 ($p < 0.001$) β = 0.3 ($p < 0.001$) VIF = 5.02
Third year requests	NA	NA	coef = 4.96 ($p < 0.001$) SE = 0.39 ($p < 0.001$) β = 0.47 ($p < 0.001$) VIF = 4.96
Year of release	coef = -40.3 ($p < 0.001$) SE = 1.76 ($p < 0.001$) β = -0.62 ($p < 0.001$) VIF = 1.1	coef = -8.85 ($p < 0.001$) SE = 1.85 ($p < 0.001$) β = -0.14 ($p < 0.001$) VIF = 2.31	coef = -9.56 ($p < 0.001$) SE = 1.65 ($p < 0.001$) β = -0.14 ($p < 0.001$) VIF = 2.32
R-squared	0.599 ($p < 0.001$)	0.7895 ($p < 0.001$)	0.833 ($p < 0.001$)

This analysis shows that early requests are good predictors for how many requests datasets will go on to receive in later years. In fact, the two- and three-year models perform almost as well in predicting later requests as they do in predicting total requests. This finding provides further evidence that requests early in a dataset's life can be helpful in predicting patterns of long-term reuse among dbGaP datasets.

5.1.2 NHLBI Results

The NHLBI analysis includes 143 datasets with a total of 3,860 requests between 2000 and 2017. Figure 5-3 shows the number of requests datasets in each decile received in each year since their release (not cumulative requests). Table 5-5 shows the range and distribution of NHLBI datasets within deciles.

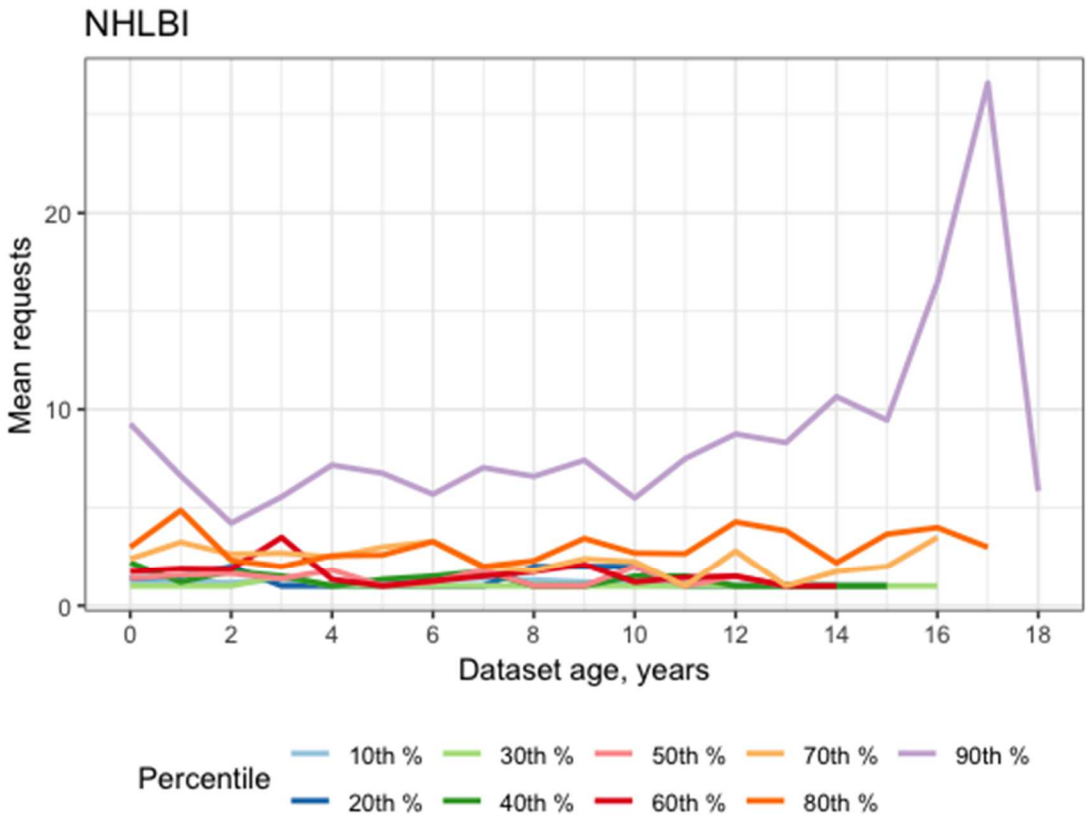

Figure 5-3. Mean requests by year for NHLBI datasets in each decile, by age of the dataset at time of request.

Table 5-5. Distribution of NHLBI datasets by request deciles for requests made between 2000 and 2017.

Decile	Request count range	Number of datasets in percentile	Mean age of datasets, years
10th percentile	2 or fewer	39	4.45 (range 0 – 17)
20th percentile	3	19	6.74 (range = 1 – 15)
30th percentile	4	22	8.64 (range = 3 – 17)
40th percentile	5	46	9.04 (range = 0 – 15)
50th percentile	6 – 8	70	6.73 (range = 2 – 12)
60th percentile	9 – 14	97	9.7 (range = 2 – 15)
70th percentile	15 – 21	107	9.49 (range = 2 – 16)
80th percentile	22 – 34	127	11.83 (range = 2 – 16)
90th percentile	35 or more	367	13.26 (range = 1 – 17)

Figure 5-3 reveals a markedly different pattern of requests from what was observed within the dbGaP analysis. NHLBI requests appear to follow no demonstrable pattern of requests at all. Part of the variability seen in Figure 5-3 is due to the sparsity of datasets that reach age 17 or 18. For example, given that only eight datasets exist that had been around for 17 years by 2017, the notable spike in requests in year 17 is probably not a meaningful finding; with so few datasets to consider in that age range, one or two outliers will have more of an effect on the mean than in a larger pool of datasets.

A further consideration in this analysis is that methods for requesting and accessing the data changed significantly over the course of the nearly twenty years considered in the full 2000 – 2017 analysis. In September 2009, NHLBI launched the BioLINCC website as it exists in its current form, which permits users to request and access data through a secure web portal (Giffen et al., 2015). Prior to the launch of the site, requestors were required to submit a paper request form by mail, and datasets were disseminated to approved requestors by mailing them a CD-ROM within two weeks (National Heart, Lung, and Blood Institute, 2008). Given the very different means of accessing data before and after the launch of the BioLINCC site, it seems reasonable to expect that patterns of requests from the two periods would likely differ.

To determine whether request patterns were more predictable after the launch of the BioLINCC site, I repeated this analysis including only the 90 datasets released between 2010 (the first complete year that BioLINCC was online) and 2017, and the

3,704 requests those datasets received during that time. However, as Figure 5-4 demonstrates, this subset shows no more coherent pattern than did the entire set.

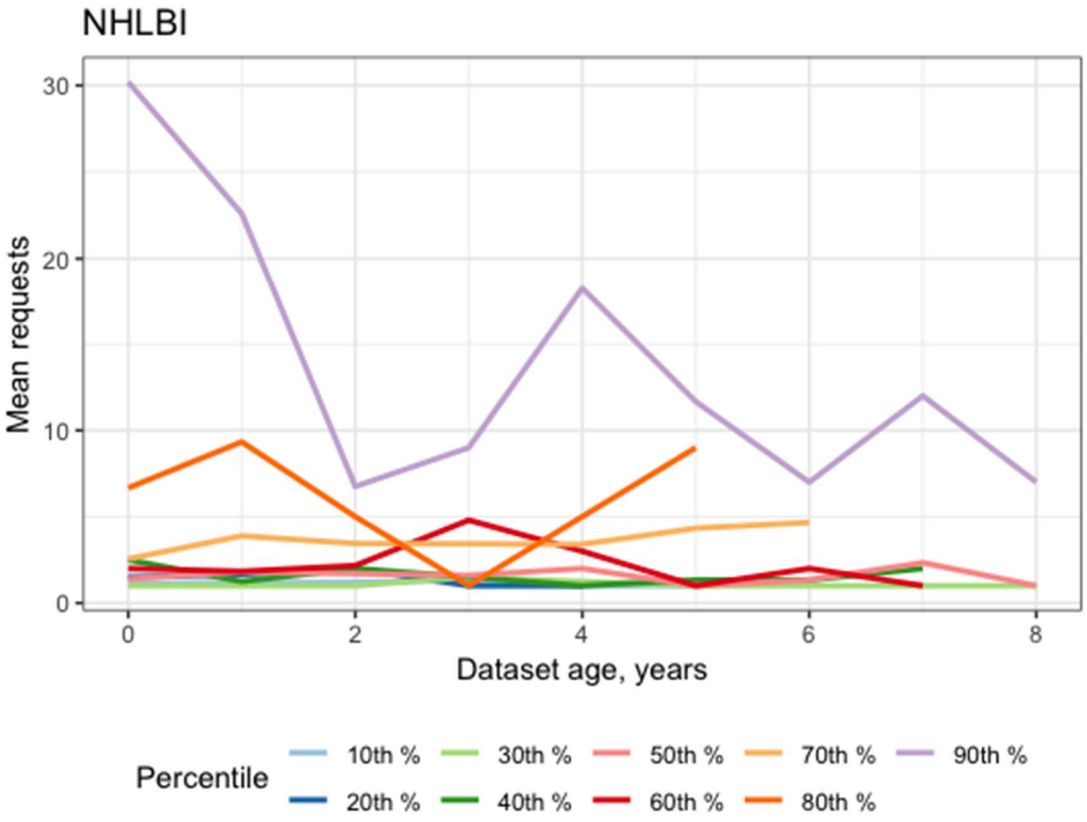

Figure 5-4. Mean requests by year for NHLBI datasets released between 2009 and 2017 in each decile, by age of the dataset at time of request.

While it seems apparent that dbGaP requests are likely a cumulative advantage process, this analysis suggests that the same may not true for NHLBI requests. However, the seeming randomness of the NHLBI data may be based more on the relative sparsity of this set compared to dbGaP's. With 982 datasets to NHLBI's 96, and a whopping 100,115 requests to NHLBI's 3,704, the dbGaP data massively dwarfs the NHLBI data. While it's possible that NHLBI request patterns

over time are significantly different from dbGaP's, and in fact seem to follow no real pattern at all, it seems just as likely that this is simply too small a set to yield meaningful findings.

Nonetheless, I did proceed with regression analysis of the NHLBI 2010 – 2017 release subset. Total requests were strongly correlated with first- and second-year requests (correlation coefficient = 0.9 for both) and but only moderately correlated with third-year requests (correlation coefficient = 0.4). Total number of requests is weakly positively correlated with calendar year of the dataset's release (correlation coefficient = 0.1), suggesting that older datasets are slightly likely to have fewer requests.

Table 5-6 summarizes results from these three regression models fit to the 2010 – 2017 request data. All models are statistically significant at $p < 0.001$.

Table 5-6. Results of regression analysis showing effects of requests during years one, two, and three of a NHLBI dataset's life on the total number of requests during the 2010 – 2017 period.

	One-Year Model	**Two-Year Model**	**Three-Year Model**
First year requests	coef = 1.63 ($p < 0.001$) SE = 0.13 ($p < 0.001$) β = 0.9 ($p < 0.001$) VIF = 1.03	coef = -0.66 ($p > 0.05$) SE = 0.56 ($p > 0.05$) β = 0.38 ($p > 0.05$) VIF = 17.44	coef = 0.52 ($p > 0.05$) SE = 0.45 ($p > 0.05$) β = 0.3 ($p > 0.05$) VIF = 23.29
Second year requests	NA	coef = 3.81 ($p < 0.001$) SE = 0.91 ($p < 0.001$) β = 1.33 ($p < 0.001$) VIF = 17.27	coef = 1.82 ($p < 0.05$) SE = 0.75 ($p < 0.05$) β = 0.64 ($p < 0.05$) VIF = 23.22

	One-Year Model	Two-Year Model	Three-Year Model
Third year requests	NA	NA	coef = 5 (p < 0.001) SE = 0.88 (p < 0.001) β = 0.31 (p < 0.001) VIF = 1.38
Year of release	coef = -4.15(p = 0.01) SE = 1.37 (p = 0.01) β = -0.22 (p = 0.01) VIF = 1.03	coef = -5.07 (p = 0.01) SE = 1.62 (p =0.01) β = -0.21 (p = 0.01) VIF = 1.04	coef = -5.57 (p < 0.001) SE = 1.34 (p < 0.001) β = -0.21 (p < 0.001) VIF = 1.06
R-squared	0.795 (p < 0.001)	0.888 (p < 0.001)	0.955 (p < 0.001)

The three-year model is the best fit, accounting for more than 95% of the variability in long-term requests. However, while the R-squared values for the NHLBI regression models are higher than the models for the corresponding years of requests in dbGaP, the Beta values are higher for the NHLBI models, suggesting that the year of release has a more significant impact on NHLBI requests. That is, part of the better predictive power in the NHLBI models is simply due to the fact that older datasets have had more time to accrue requests, and not because requests within the first several years are more highly predictive within NHLBI than within dbGaP. In addition, the two- and three-year models' VIF values indicate that there is a level of collinearity between first and second year requests. However, this finding is not necessarily problematic since its mechanism can likely be understood by the fact that datasets receive similar amounts of requests in their first and second years. Moreover,

first- and second-year requests are not perfectly collinear in the sense that one predicts the other in the way that variables like age and date of birth do. Finally, while methods exist to address collinearity, they do not perform much better than standard regression models, and many statisticians recommend simply ignoring collinearity (Dormann et al., 2013).

I also considered the role of first, second, and third year requests in predicting later requests, made in the fourth year and beyond. This analysis includes the 49 datasets that were released before 2015 (that is, those that were old enough to have more than three years' worth of requests). All three models also include release year to control for the influence of a dataset's age on the number of requests. All three models are statistically significant, although the one-year model achieved a higher (but still significant) *p*-value. Table 5-7 shows a summary of the results.

Table 5-7. Results of regression analysis showing effects of requests during year one, two, and three of an NHLBI dataset's life on the total number of requests in the fourth year and later during the 2009 – 2017 period.

	One-Year Model	**Two-Year Model**	**Three-Year Model**
First year requests	coef = 4.6 (p = 0.03) SE = 2.1 (p = 0.03) β = 0.3 (p = 0.03) VIF = 1.13	coef = 1.43 (p = 0.3) SE = 1.48 (p = 0.3) β = 0.09 (p = 0.3) VIF = 1.24	coef = 0.12 (p = 0.9) SE = 1.48 (p = 0.9) β = 0.008 (p = 0.9) VIF = 1.39
Second year requests	NA	coef = 5.13 (p < 0.001) SE = 0.68 (p < 0.001) β = 0.69 (p < 0.001) VIF = 1.1	coef = 3.8 (p < 0.001) SE = 0.82 (p < 0.001) β = 0.51 (p < 0.001) VIF = 1.79

	One-Year Model	Two-Year Model	Three-Year Model
Third year requests	NA	NA	coef = 2.31 (p = 0.01) SE = 0.88 (p = 0.01) β = 0.3 (p = 0.01) VIF = 1.99
Year of release	coef = -4.55 (p < 0.001) SE = 1.26 (p < 0.001) β = -0.5 (p < 0.001) VIF = 1.13	coef = -4.61 (p < 0.001) SE = 0.85 (p < 0.001) β = -0.5 (p < 0.001) VIF = 1.13	coef = -4.25 (p < 0.001) SE = 0.81 (p < 0.001) β = -0.46 (p < 0.001) VIF = 1.17
R-squared	0.2332 (p = 0.002)	0.6593 (p < 0.001)	0.7057 (p < 0.001)

Unlike with the dbGaP data, first year requests do not appear to be a good predictor for reuse later in an NHLBI dataset's life. In fact, while the one-year model did achieve statistical significance, the first-year request variable barely did (p = 0.03), and first-year requests did not achieve statistical significant in the two- and three-year models. Further, the Beta value for first-year requests was lower than for year of release in each model; the very low Beta values in the two- and three-year models indicate that first-year requests have very little impact at all on later requests.

Because first year requests are such a poor predictor for requests later in the dataset's life, I also fit models for second-year requests only and second- and third-year requests. Table 5-8 summarizes these models.

Table 5-8. Results of regression analysis showing effects of requests during year two and three of an NHLBI dataset's life on the total number of requests in the fourth year and later during the 2009 – 2017 period.

	Second Year Only Model	**Second and Third Year Model**
Second year requests	coef = 5.32 ($p < 0.001$) SE = 0.66 ($p < 0.001$) β = 0.71 ($p < 0.001$) VIF = 1.01	coef = 3.8 ($p < 0.001$) SE = 0.81 ($p < 0.001$) β = 0.51 ($p < 0.001$) VIF = 1.79
Third year requests	NA	coef = 2.33 ($p = 0.006$) SE = 0.82 ($p = 0.006$) β = 0.31 ($p = 0.006$) VIF = 1.77
Year of release	coef = -4.34 ($p < 0.001$) SE = 0.8 ($p < 0.001$) β = -0.47 ($p < 0.001$) VIF = 1.01	coef = -4.23 ($p < 0.001$) SE = 0.75 ($p < 0.001$) β = -0.46 ($p < 0.001$) VIF = 1.02
R-squared	0.6522 ($p < 0.001$)	0.7056 ($p < 0.001$)

Interestingly, these models performed substantially better than the models that include the first year. For example, considering first-year requests only accounts for only 23% of the variability in later requests, while considering second-year requests only accounts for 65% of the variability. The second and third year model likewise performs better than the first and second year model. This finding suggests that, while first-year requests are a good predictor for dbGaP dataset's long-term reuse, the same is not true for NHLBI. Rather, it appears that it takes longer for a dataset in NHLBI to be "noticed" and start receiving requests. This finding is further supported by the fact that while 27% of dbGaP datasets received no requests in their first year, 40% of the NHLBI datasets received no requests in their first year. A Welch unpaired two-sample t-test shows that the means of the maximum semantic similarity scores for

dbGaP and NIDDK request/dataset pairs are significantly different ($t = 2.49$, $df = 83.34$, $p = 0.01$).

5.1.3 Summary of Findings

Here, I considered whether there are temporal patterns to dataset requests by considering the number of requests datasets receive by year since their release, rather than calendar year. Specifically, I tested the hypothesis that patterns of requests relative to the original dataset release date will be similar to patterns of citations to articles relative to their publication date. The findings of the dbGaP analysis support this hypothesis; like citations to articles, requests to dbGaP datasets appear to be a cumulative advantage process, with highly requested datasets going on to receive even more requests over time. Except for the tier of most- requested datasets, dbGaP datasets requests peak around the third year after a dataset is released and gradually decline over time, a pattern again seen in citations to articles (Parolo et al., 2015). Finally, regardless of when a dataset was released, the number of requests it receives in the first few years of its life are a good prediction for how many requests it will go on to receive.

On the other hand, this hypothesis did not hold true with the NHLBI analysis. While mean requests by decile followed a clean pattern in the dbGaP datasets, requests in the NHLBI datasets appeared to follow virtually no pattern at all. Even when considering only the subset of datasets that had been released during the time that the request process existed in electronic form, no pattern appeared to exist.

However, interestingly, these analyses found that first-year requests, which were a good predictor of later requests in the dbGaP datasets, were actually a very poor predictor for later requests in NHLBI. Instead, second- and third-year requests showed good predictive power, suggesting that patterns of reuse differ between NHLBI and dbGaP. NHLBI datasets take longer to begin accruing requests, a finding that could suggest genomic data starts being requested earlier than clinical data, but these differences could also be due to characteristics of the repositories themselves. For example, perhaps dbGaP does more to raise awareness of its datasets among the community of researchers who use them. If a dataset's release is not publicized, researchers would likely not be aware of its existence until later after its release, perhaps when articles start to cite the dataset (which would likely correspond with the second and third year after the dataset's release). Further research into how these repositories promote outreach to their research communities and how researchers typically find datasets to reuse could help explain these findings.

5.3 Research Question 4: Are there dataset topics that are more highly requested?

The datasets contained within the three repositories considered here vary, in some cases significantly, in the number of requests they have received. The length of time a dataset has been available likely plays some role in accounting for more requests; a dataset released in 2009 has had more time to be requested than a dataset released in 2017, so it stands to reason that the total number of requests accrued by age would differ. However, as was demonstrated in Section 5.1, the variations in

request numbers can only partly be explained by how long a dataset has been available.

The repositories considered here contain datasets that cover a wide range of different conditions and disorders, from the very common (such as heart disease, which is the leading cause of death in the US) to the very rare (such as biliary atresia, a rare liver disorder affecting about 1 in 20,000 live births in the US), as well as including some reference sets of healthy human subjects (Hopkins, Yazigi, & Nylund, 2017; National Center for Health Statistics, 2017). Given that the burden of disease for various conditions differs widely, biomedical research funders and pharmaceutical development companies understandably tend to focus more money and effort on certain diseases. Likewise, it seems logical that some of the topics within the datasets from these three repositories would be more "popular" than others and would therefore receive more requests. Here I aim to explore whether variations in numbers of requests are due to the subject coverage of the dataset – in other words, are some topics more highly requested?

5.2.1 Defining Topics

Determining whether some topics are more requested requires first defining what the major topics in repositories are, a task that is not as straightforward as it might seem. Of course, datasets could simply be categorized into topics based on the primary condition the dataset covers, but there are other characteristics of datasets that might influence requests. For example, some of the studies in the NIDDK and

NHLBI repositories are longitudinal, following participants over the course of decades. Such datasets provide a wealth of data that could be useful for a range of research purposes, regardless of the topic focus of the original study. As the semantic similarity analysis described in Section 4.1.2 demonstrated, datasets often end up being reused in the context of topics that differ significantly from the topic for which the dataset was originally requested, so at least in some cases, the topic of the dataset is not what makes it appealing for reusers.

Rather than make my own assumptions about how to divide datasets into topics, I utilized a topic modeling approach that used a technique called latent Dirichlet allocation (LDA) to sort datasets into topics with other datasets that were most similar to them. For this analysis, I considered each repository separately, since the subject coverage and request patterns for the three repositories are very different, and used the descriptions of each dataset as the corpus for text mining. I wrote custom R scripts that retrieved the descriptions for each dataset from the three repositories and removed extraneous text, such as HTML tags and section headers, leaving only the dataset description. Next, I removed common English language stopwords (such as "and" and "the"), as well as a set of custom stopwords, which were terms such as "patients" or "research" that appeared in many of the dataset descriptions and did not provide meaningful context (the full list of stop word is included as Appendix B).

Next, I experimented with several text processing techniques to determine which processes would yield the most meaningful input for the topic-modeling algorithm. First, various forms of the same word needed to be processed to arrive at a

common form that would prevent the algorithm from considering them as two different words. For example, the words “decrease,” “decreases,” “decreased,” and “decreasing” are all forms of the common stem “decrease,” and therefore should be considered as one, rather than four separate terms. I tested the stemming algorithm in the R package SnowballC (Bouchet-Valat, 2019), but found it to be aggressive in stemming words, essentially indiscriminately stripping many “-s”, “-d”, and “-ing” endings from words that should not have been stemmed, resulting in many terms being reduced to the same stem when in fact they were not topically similar. Instead, I used lemmatization, a process that is more computationally intensive but achieved better results by identifying the term within a pre-defined dictionary and thereby determining the correct root word (Rinker, 2018).

Once terms had been lemmatized, I experimented to determine the unit of analysis that would provide the most meaningful input for the LDA algorithm. I started with using individual lemmatized words, producing a count of the number of times each word appeared in a given description. This approach is referred to as a “bag of words” approach, since it simply counts words in the text without consideration for the context of the word within the description. Given the complexity of the concepts within the descriptions of the datasets, the bag of words approach was not effective. For example, consider the topics “elevated blood pressure,” “elevated risk,” “elevated glucose levels,” and “elevated liver enzymes.” These four terms refer to very different concepts, but the bag of words approach would consider them similar because they all contain the term “elevated.”

Instead, I experimented with bigrams and trigrams, sets of two or three words appearing in conjunction with each other. This approach provides greater consideration of the words in context. For example, “elevated blood pressure” would be split into bigrams “elevated blood” and “blood pressure.” Some less meaningful connections would still be made with the “elevated blood” bigram, connecting this description to ones referring for example, to “elevated blood glucose” or “elevated blood platelets.” However, it would also have the bigram of “blood pressure” to connect it to concepts such as “high blood pressure” and “blood pressure measurement.” I processed all descriptions into both bigrams and trigrams and found that bigrams provided the most useful sets of terms for these descriptions.

The lemmatized bigrams were then arranged in a document-term-matrix (dtm), which gives a count for the number of times a bigram appears in each dataset description. The dtm serves as the input for the LDA algorithm, which sorts the documents (i.e. the dataset descriptions) into k groups of similar documents, where k is the user-defined number of topics. To some extent, determining the number of topics is a matter of trial and error, experimenting with different values of k until meaningful topics appear. However, the R package ldatuning provides some metrics to assist in determining the optimal value of k. For each repository, I tested for all values of k between 5 and 25 to determine the optimal number of groups. The package returns results of four metrics, two for which the optimal value should be as low as possible and two for which the optimal value should be as high as possible.

Figure 5-5, Figure 5-6, and Figure 5-7 show output from the ldatuning package, run for *k* between 5 and 25 for each of the repositories.

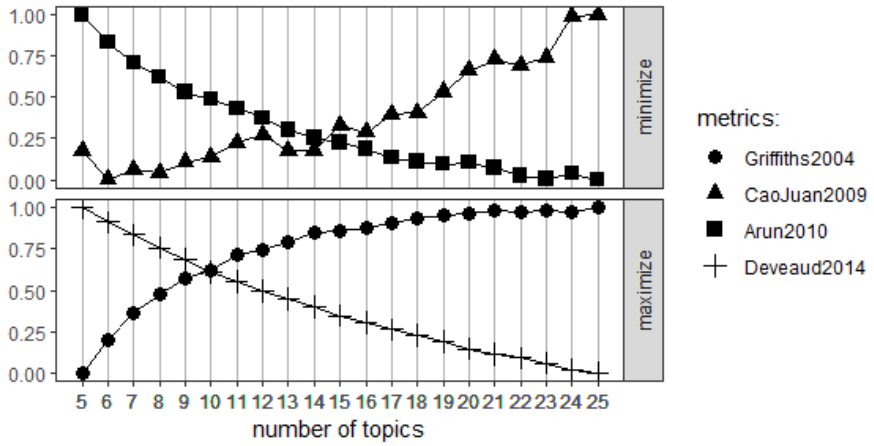

Figure 5-5. Output from ldatuning package for the dbGaP dataset descriptions.

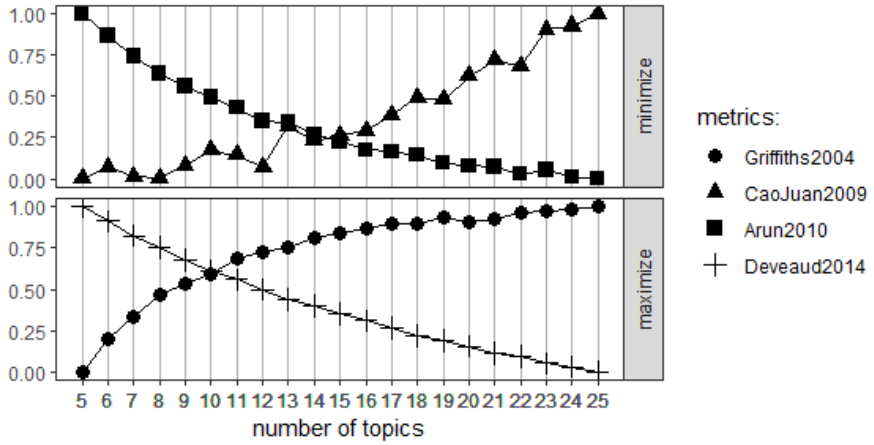

Figure 5-6. Output from ldatuning package for the NHLBI dataset descriptions.

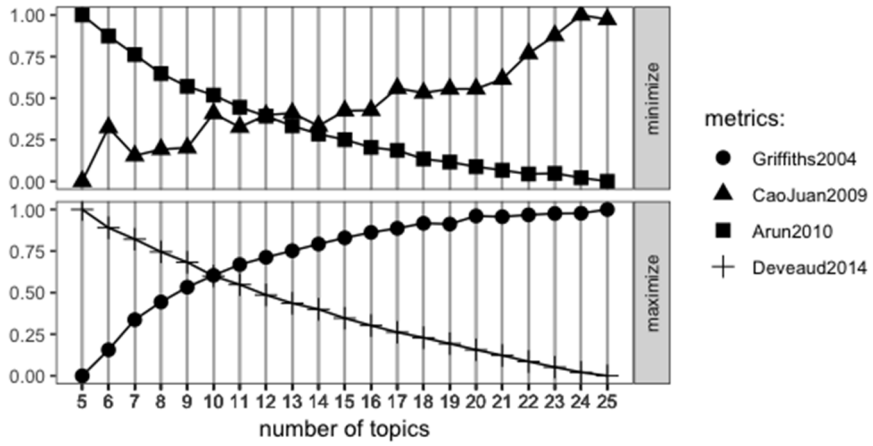

Figure 5-7. Output from the ldatuning package for the NIDDK dataset descriptions.

Based on this output, I ran the LDA algorithm with the *k* values that appeared optimal, comparing a few variations where the ldatuning package showed some ambiguity. For example, for NIDDK, I tried models with *k* of 13 and 14, and examined the outputs to determine which model provided the most meaningful groupings. Each term was assigned a beta value that indicated how strongly it was associated with a topic. Reviewing the ten terms with the highest beta for each grouping helped provide insight as to whether the grouping was meaningful and what the topic was about. Figure 5-8 shows an example of a chart showing the top ten terms in topic 7 of the 14-group NIDDK model. Appendix C contains the full set of charts for all topics for each repository.

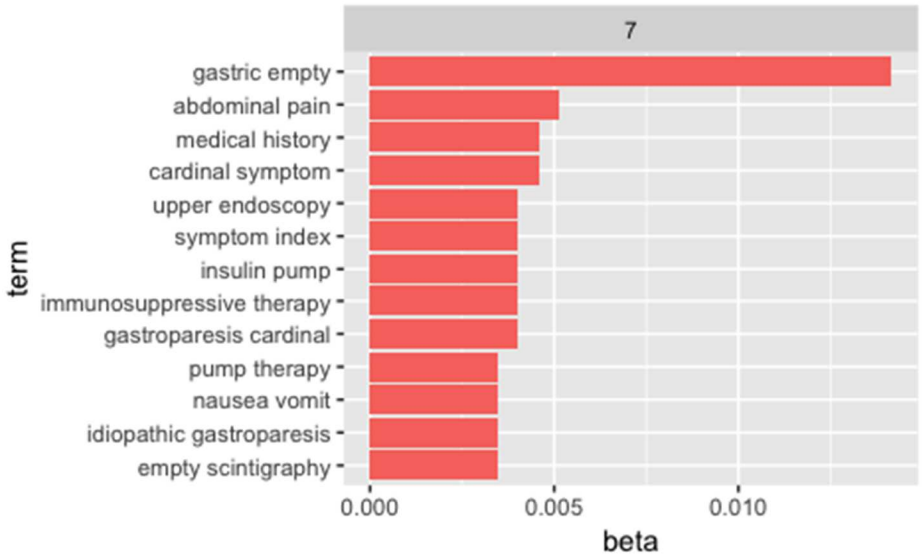

Figure 5-8. An example of a chart showing the top ten terms in topic 7 of the 14-group NIDDK model with its corresponding beta value.

A consideration of the terms shown in Figure 5-8 confirms that this is a logical grouping of documents and also provides insight into how the topic can be described. Gastroparesis is a disorder characterized by delayed gastric emptying, resulting in nausea, vomiting, and abdominal pain. It is diagnosed by a gastric emptying scintigraphy test, as well as upper endoscopy, and the severity of the condition can be quantified using the Gastroparesis Cardinal Symptom Index. Gastroparesis often occurs in insulin-dependent patients with diabetes, but in non-diabetics, it is known as idiopathic gastroparesis. Based on this list of terms, this topic appears to contain datasets that are about gastroparesis. Of course, some of the terms are general enough that they might not refer to gastroparesis – for example, nausea and vomiting are common to many illnesses, and upper endoscopy is used in the context of many gastrointestinal disorders. Therefore, I also reviewed the list of

datasets that had been classified as belonging to topic 7 to determine whether "gastroparesis" was an accurate title for this grouping, as well as to ensure that the grouping of datasets also made logical sense. In reviewing the datasets in this topic, gastroparesis was the most common topic, but a few datasets also contained data about other rare gastrointestinal disorders, so I expanded the title of this topic to "gastroparesis and other GI disorders."

I conducted this procedure with the descriptions of datasets from each of the repositories. For both NIDDK and NHLBI, the optimal number of groupings was 14, the topics of which are described below. However, the LDA algorithm was less useful in analyzing the dbGaP dataset descriptions. The tuning algorithms suggested using a *k* of 11, which yielded groupings of datasets that seemed mostly unconnected and for which I could not find meaningful topic descriptions. I experimented with a range of different values for *k*, but was not able to obtain groupings that made sense. The success of the topic modeling in NIDDK and NHLBI might be due to the fact that the datasets in these repositories do generally fall within a relatively constrained range of topics – after all, they only collect datasets related to diabetes, digestive disorders, and kidney diseases (NIDDK) and heart-, lung-, and blood-disorders (NHLBI). dbGaP, by comparison, contains thousands of datasets spanning the range of human disease and health, so it may be that the range of topics is too complex to be meaningfully captured by the LDA algorithm. Of course, it is also possible that the groupings the algorithm made did actually have some meaning, but it was too obscure

for me to understand (such as, “datasets with a principal investigator named Jim”) and that also would have been unlikely to provide a meaningful basis for this analysis.

Since the LDA algorithm was ineffective for the dbGaP datasets, I instead categorized them based on the “primary phenotype” (that is, the main disease or characteristic of interest in the dataset) reported on the dbGaP website for each dataset. The 1,150 datasets had 452 unique primary phenotypes; to achieve a more manageable number of topics, I grouped the datasets into 18 broad topics as described below, using the MeSH trees into which each phenotype term fell as a guide. Because the dbGaP dataset also contains a large number of datasets covering different types of cancers, I also further categorized cancer datasets with the type of cancer they described and conducted a sub-analysis of these datasets.

5.2.2 Comparing Requests Across Topics

Because datasets were not evenly distributed among the topics, raw request counts would not provide a fair comparison for considering request rates. For example, consider the top two most requested topics in the NIDDK repository, Chronic Kidney Diseases and Type 2 and Gestational Diabetes (shortened here for convenience to CKD and T2D), which have received 125 and 104 requests and account for 32% and 27%, respectively, of all requests submitted to the NIDDK repository. However, there are almost twice as many CKD datasets (13, or 14% of all NIDDK datasets) as there are T2D datasets (7, or 8% of all NIDDK datasets). Even though the request counts are similar, the 104 requests for T2D topics are spread

among a much smaller set of datasets, and therefore cannot reasonably be compared to the CKD requests.

To account for the differences in number of datasets per topic, I calculated a request to dataset (RTD) ratio. First, I calculated the proportion of requests by dividing the number of requests in a topic by the total number of requests in the repository. Similarly, I calculated the proportion of datasets by dividing the number of datasets in a topic by the total number of datasets in the repository. Dividing the proportion of requests by the proportion of datasets, I arrived at the request ratio. Figure 5-9 provides a visual explanation of this process. In this example, topic A's request ratio is calculated by dividing the proportion of its requests (70 requests for topic A datasets divided by 192 total requests for datasets in the repository = 0.36) by the proportion of datasets in the topic (4 datasets in topic A divided by 6 datasets total in the repository = 0.67), arriving at a ratio of 0.54.

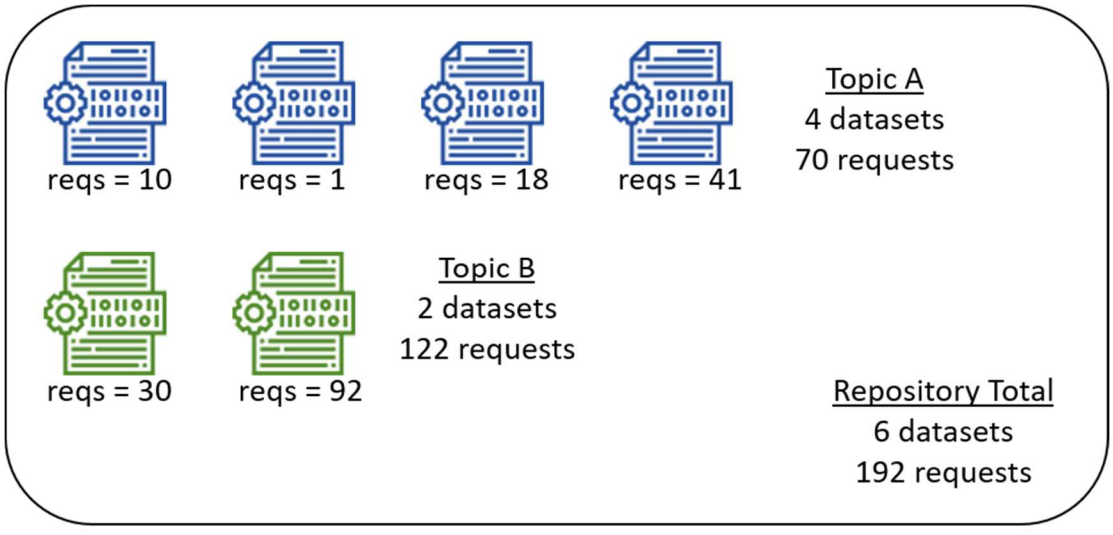

Figure 5-9. Visual explanation of request ratio calculation.

A ratio of 1 would indicate that a topic received as many requests as would be expected based on the number of datasets in the topic. If every topic in a repository received a score of 1, it would mean that every topic had been requested at the same rate, and essentially all topics were equally popular. A topic with a ratio of greater than 1 is over-requested based on how many datasets it contains; for example, a ratio of 2 would mean the topic had received twice as many requests as would be expected based on the number of datasets it contained. Similarly, a ratio of less than 1 meant the topic was under-requested; a ratio of 0.3 would mean the topic had received only 30% as many requests as would be expected based on the number of datasets it contained.

To revisit our NIDDK example, the CKD topic has a request proportion of 0.323 and a dataset proportion of 0.141, yielding a request ratio of 2.29. The T2D topic has a request proportion of 0.269 and a dataset ratio of 0.076, yielding a request ratio of 3.54. Both topics are highly requested; with a request ratio of greater than 1, they received more requests than would be expected if all topics were requested equally. However, despite the T2D topic having received 21 fewer requests than the CKD topic, it has actually outperformed CKD by 1.5 times, given that T2D had fewer datasets than CKD overall.

In addition to considering the total RTD ratio for each topic, I also calculated a yearly RTD ratio to explore whether topic popularity remained consistent or whether some topics gained or lost popularity over time. To calculate the yearly RTD ratio, I used annual request counts and cumulative rather than total dataset counts. For

example, in 2009 dbGaP contained eight datasets in the Cancer topic, which received a total of 230 requests in that year. By 2010, an additional 10 datasets had been added for a total of 18 datasets that received 592 requests in the Cancer topic. To calculate the 2009 RTD ratio for the Cancer topic, I used the proportions for that year; the eight datasets that were about Cancer constituted 4.5% of the 178 datasets that existed in dbGaP at that time. By 2010, dbGaP contained 243 total datasets, so the 18 Cancer datasets now constituted 7.4% of the total.

5.2.3 dbGaP Results

1,133 datasets from dbGaP were sorted into 18 topics based on their primary phenotype. These datasets had received a total of 104,337 requests between 2008 and 2018. Table 5-9 shows the distribution of datasets and requests among the 18 topics and each topic's RTD ratio.

Table 5-9. Distribution of dbGaP datasets and requests among 18 topics derived from the assigned primary phenotype, and calculated request to dataset (RTD) ratio.

Topic	Datasets	Requests	RTD Ratio
Blood and Cardiovascular	269 (23.7%)	66,725 (64%)	2.69
Mental Disorders	39 (3.4%)	3,117 (3%)	0.87
Eye Disorders	20 (1.8%)	1,298 (1.2%)	0.7
Normal	48 (4.2%)	2,668 (2.6%)	0.6
Women's Health and Pregnancy	26 (2.3%)	1,412 (1.4%)	0.59
Cancer	319 (28.2%)	17,208 (16.5%)	0.59
Neurological	86 (7.6%)	4,154 (4%)	0.52
Lung and Respiratory Disorders	38 (3.4%)	1,653 (1.6%)	0.47
Substance Use Disorders	18 (1.6%)	729 (0.7%)	0.44
Metabolic Diseases	57 (5%)	2,147 (2.1%)	0.41
Skin Disorders	7 (0.6%)	238 (0.2%)	0.37
Other	39 (3.4%)	926 (0.9%)	0.26

Topic	Datasets	Requests	RTD Ratio
Musculoskeletal	17 (1.5%)	293 (0.28%)	0.19
GI Disorders	26 (2.3%)	365 (0.3%)	0.15
Congenital Disorders	70 (6.2%)	910 (0.9%)	0.14
Immune and Autoimmune Disorders	16 (1.4%)	173 (0.2%)	0.12
Urogenital Disorders	18 (1.6%)	190 (0.2%)	0.11
Infectious Disease	20 (1.8%)	131 (0.1%)	0.07

The mean RTD ratio for dbGaP topics is 0.52, indicating that disparity exists among the various topics. Most of this disparity comes from the Blood and Cardiovascular topic being highly over-requested, receiving requests at a rate nearly triple would be expected based on the number of datasets in the category. Six categories also have ratios of less than 0.2, having received less than 20% as many requests as would be expected.

Figure 5-10 shows the annual RTD ratios for each topic between 2008 and 2018. The dashed line indicates a ratio of 1; values below the line indicate higher-than-expected requests, and values below the line, lower-than-expected requests. Annual results are similar to the overall results described above, and RTD ratios remain generally steady for most topics over time. However, a few topics do show change over time. Blood and Cardiovascular datasets, already over-requested in 2008 with a ratio of 1.14, continues to rise in popularity, eventually reaching a high RTD of 2.3 in 2017. Conversely, datasets in the Mental Disorders topic see their ratio decline over time; with RTD ratios over 1 and even approaching 2 in most years between 2008 and 2012, the ratio declined to just over 0.6 by 2018.

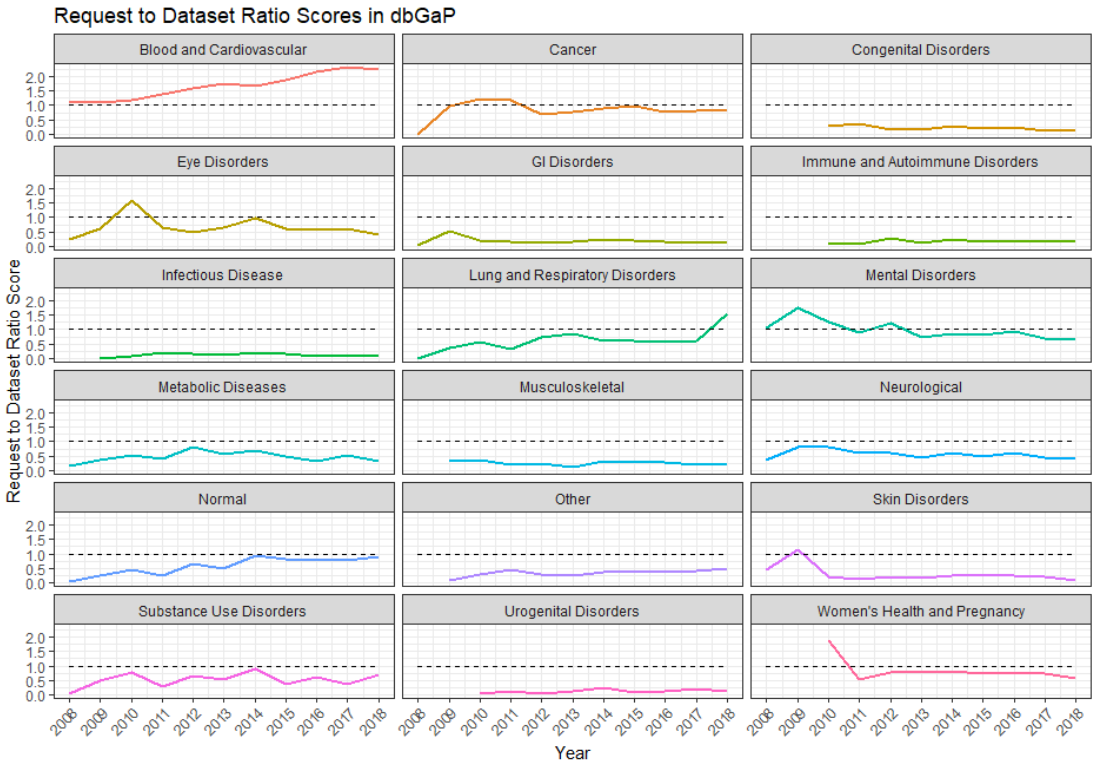

Figure 5-10. Request to dataset ratios for dbGaP datasets, by topic, calculated annually from 2008 – 2018.

In addition to considering the full dbGaP repository, I also performed this analysis for the 319 datasets that contained data about cancer to determine whether differences existed in requests for data about specific types of cancer. These datasets received 17,208 requests between 2008 and 2018. I classified them into ten groups based on primary cancer site, with one group for datasets that included multiple types of cancer as well as forms of cancer that could not be categorized into one of the other nine types. Table 5-10 shows the distribution of datasets and requests among the ten cancer types and each topic's RTD ratio.

Table 5-10. Distribution of dbGaP datasets specific to cancer and their requests among 10 cancer topics derived from the assigned primary phenotype, and calculated request to dataset (RTD) ratio.

Topic	Datasets	Requests	RTD Ratio
Other or Multiple Cancers	39 (12.2%)	4,388 (25.5%)	2.09
Blood Cancer	63 (19.7%)	3,950 (23%)	1.16
Bone and Soft Tissue Cancers	11 (3.4%)	618 (3.6%)	1.04
Urogenital Cancers	32 (10%)	1,581 (9.2%)	0.92
Prostate Cancer	28 (8.8%)	1,373 (8%)	0.91
Lung Cancer	23 (7.2%)	1,116 (6.5%)	0.9
Brain and Nervous System Cancers	21 (6.6%)	965 (5.6%)	0.85
Breast Cancer	36 (11.3%)	1,434 (8.3%)	0.74
Skin Cancers	25 (7.8%)	716 (4.2%)	0.53
GI Cancers	41 (12.9%)	1,067 (6.2%)	0.48

The mean RTD ratio among the cancer datasets is 0.96, indicating that requests are relatively evenly distributed among the topics. The Other or Multiple Cancer type is requested at a rate more than double what would be expected based on the number of datasets, but this category is influenced by a significant outlier: the Cancer Genome Atlas (TCGA). This dataset, which contains detailed data about several different types of cancer, has been requested 2,857 times since its release in 2009, more than three times as many as the next-most requested dataset in all of dbGaP. No other dataset in dbGaP (or any of the other repositories in this study) has been requested so significantly more than TCGA; its requests alone account for 65% of requests in the Other or Multiple Cancer topic and 17% of all the requests in the subset of datasets on cancer. Without the TCGA requests, the Other or Multiple Cancer topic would only have an RTD ratio of 0.33, and if TCGA alone were considered its own topic, it would have an RTD ratio of 55.3.

Figure 5-11 shows RTD ratios for the ten cancer types for each year between 2008 and 2018, which remain mostly steady over time.

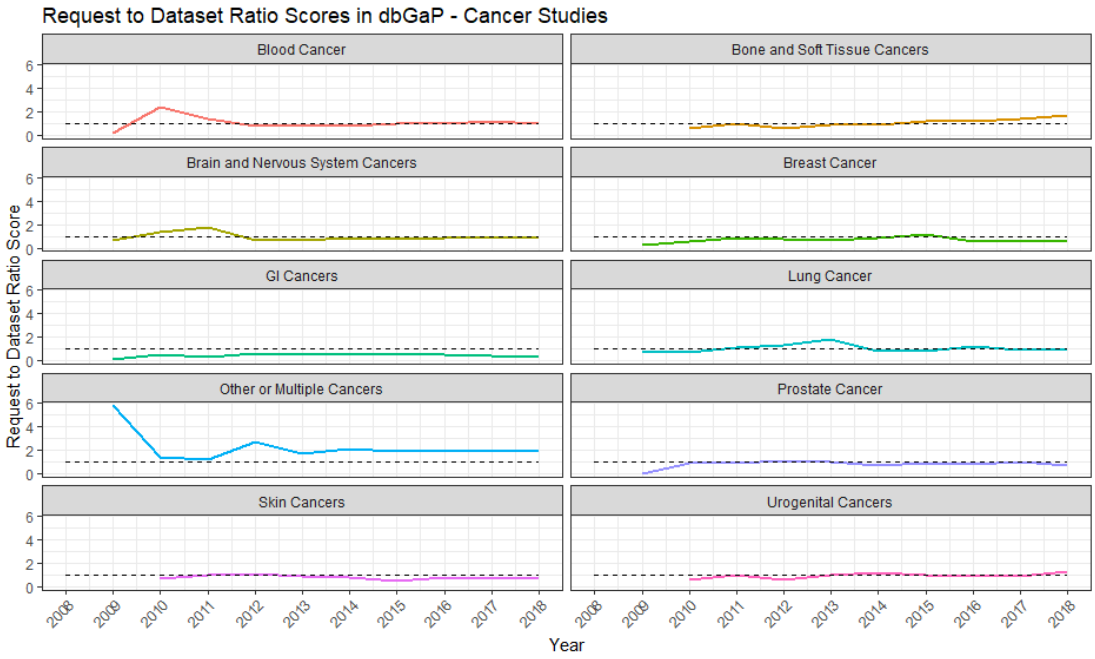

Figure 5-11. Request to dataset ratios for dbGaP datasets related to cancer, by cancer type, calculated annually from 2008 – 2018.

5.2.4 NHLBI Results

The LDA algorithm classified the 166 datasets from NHLBI into 14 different topics. These datasets had received a total of 893 requests between 2000 and 2018. Table 5-11 shows the distribution of datasets and requests among the 14 topics and each topic's RTD ratio.

Table 5-11. Distribution of NHLBI datasets and their requests among 14 topics determined by LDA, and calculated request to dataset (RTD) ratio.

Topic	Datasets	Requests	RTD Ratio
Heart Disease Treatment and Prevention	17 (10.2%)	157 (17.6%)	1.72
Lung Injuries and Mechanical	12 (7.2%)	97 (10.9%)	1.5

Topic	Datasets	Requests	RTD Ratio
Ventilation			
Population-Based Studies	7 (4.2%)	53 (5.9%)	1.41
Cardiovascular Risk Factors	11 (6.6%)	82 (9.2%)	1.39
Heart Failure and Rhythm Disorders	16 (9.6%)	106 (11.9%)	1.23
Hypertension	15 (9%)	98 (11%)	1.21
Non-Asthma Lung Diseases	14 (8.4%)	81 (9.1%)	1.08
Myocardial Ischemia	13 (7.8%)	61 (6.8%)	0.87
Sickle Cell Anemia and Blood-Borne Diseases	8 (4.8%)	32 (3.6%)	0.74
HIV and Other Viral Diseases	9 (5.4%)	33 (3.7%)	0.68
Asthma	19 (11.4%)	60 (6.7%)	0.59
Emergency Resuscitation	8 (4.8%)	17 (1.9%)	0.4
Coagulation and Sleep Disorders	5 (3%)	9 (1%)	0.33
Blood Transfusions and Marrow Transplants	12 (7.2%)	7 (0.8%)	0.11

The mean RTD ratio for all NHLBI datasets was 0.94, suggesting a relatively even distribution of requests among the 14 topics. Topics related to heart disease were particularly popular, with Heart Disease Treatment and Prevention and the related Hypertension and Cardiovascular Risk Factors topics (both of which lead to heart disease) all having RTD ratios over 1. By comparison, non-heart-related topics were more under-requested; of the seven topics with an RTD of less than 1, only one of them, Myocardial Ischemia, is related to any kind of cardiovascular disorder.

Figure 5-12 shows RTD ratio scores for each topic over time. Several of the topics do not appear across all years of this analysis; for example, the first datasets in the Coagulation and Sleep Disorders topic were not released until 2014, so that topic's first RTD ratio is recorded in that year. As with the dbGaP topics, RTD scores among the NHLBI topics remain mostly steady over time.

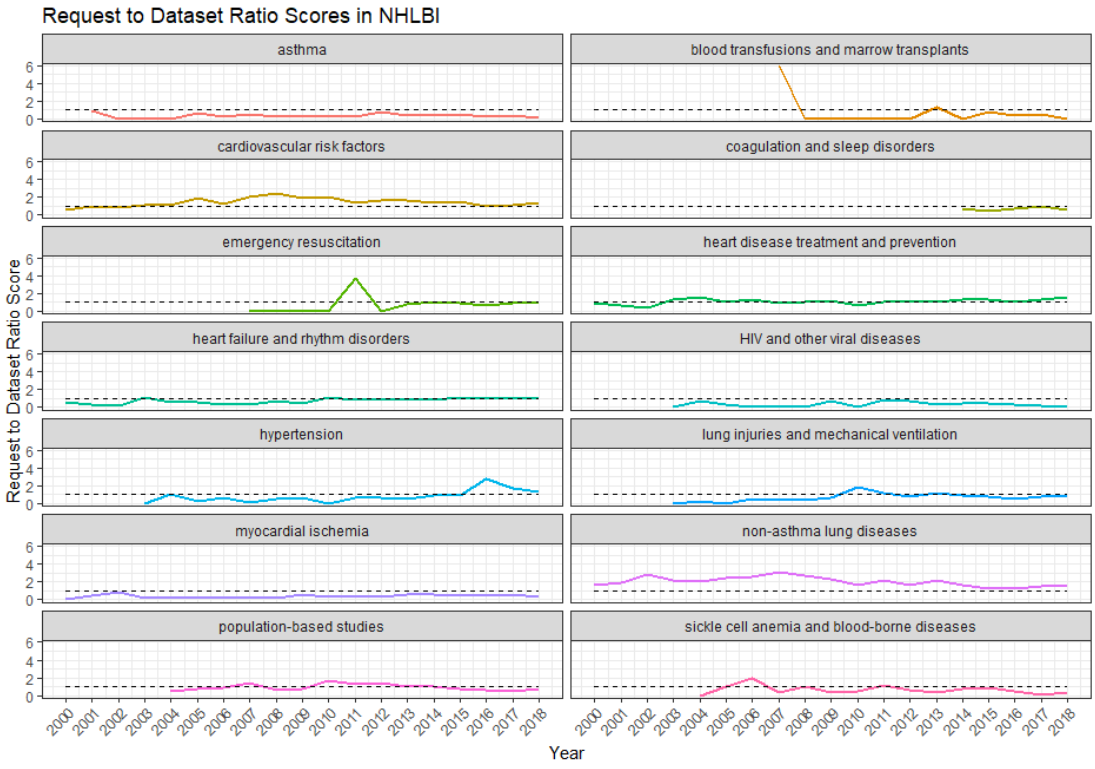

Figure 5-12. Request to dataset ratios for NHLBI datasets by topic, calculated annually from 2000 – 2018.

5.2.5 NIDDK Results

The 92 datasets in NIDDK were sorted into 14 topics determined by the LDA algorithm. These datasets received a total of 387 requests between 2013 and 2018. These datasets and requests do not represent the entire set of NIDDK datasets and requests; because the annual RTD analysis requires knowing how many total datasets existed in a given year, the date of release must be known for every dataset, but NIDDK did not track the date of release for datasets released before 2013. Because 49 of the NIDDK datasets were simply recorded as being released sometime before

2013, the earliest the annual analysis could begin was 2013. Table 5-12 shows the distribution of datasets and requests among the 14 topics and each topic's RTD ratio.

Table 5-12. Distribution of NIDDK datasets and their requests from 2013 – 2018, for 14 topics determined by LDA, and calculated request to dataset (RTD) ratio.

Topic	Datasets	Requests	RTD Ratio
Type 2 and Gestational Diabetes	7 (7.6%)	104 (26.9%)	3.53
Chronic Kidney Diseases	13 (14.1%)	125 (32.3%)	2.29
Glomerulopathies*	3 (3.2%)	11 (2.8%)	0.87
Genetics and Disease Mechanisms	8 (8.7%)	26 (6.7%)	0.77
Dialysis and Lifestyle Interventions	9 (9.8%)	27 (7%)	0.71
Nonalcoholic Liver Diseases and Bariatric surgery	9 (9.8%)	23 (5.9%)	0.61
Type 1 Diabetes	8 (8.7%)	20 (5.2%)	0.59
Hepatitis	6 (6.5%)	14 (3.6%)	0.55
Incontinence	5 (5.4%)	8 (2.1%)	0.38
Urological Disorders	12 (13%)	19 (4.9%)	0.38
Gastroparesis and GI Diseases	4 (4.3%)	6 (1.6%)	0.36
Biliary Diseases and Liver Transplantation	6 (6.5%)	4 (1%)	0.16
Islet Transplantation**	2 (2.2%)	0 (0%)	0

*diseases affecting the filtering mechanism of the kidney
**transplantation of insulin-producing cells to treat type 1 diabetes

The NIDDK topics had a mean RTD ratio of 0.86, suggesting that there is at least moderate disparity in requests among the topics. In fact, NIDDK has two of the highest RTD ratios of all three repositories, with the Type 2 and Gestational Diabetes and Chronic Kidney Diseases topics scoring 3.53 and 2.29, respectively. NIDDK is also the only repository to have a topic with an RTD score of 0, meaning the datasets in this topic have never been requested. However, only two datasets were in the Islet Transplantation topic, and both were released in 2016, so these datasets may go on to receive requests over time.

Figure 5-13 shows RTD ratio scores for each topic over time (the Islet Transplantation topic is not shown because its RTD ratio is 0). Many of the topics remain steady over time, but the yearly RTD ratios show somewhat more variability than those for NHLBI and dbGaP. For example, the Hepatitis and Dialysis and Lifestyle Interventions topics both have RTD ratios greater than one for the first two years of the analysis, but then decline and drop below 1 in 2015. The Chronic Kidney Diseases topic also sees a significant increase in its RTD ratio in 2015; while some topics in the other repositories see a spike in a single year and then drop back to the baseline rate in the following year, the increase in the RTD ratio for this topic lasts throughout this analysis.

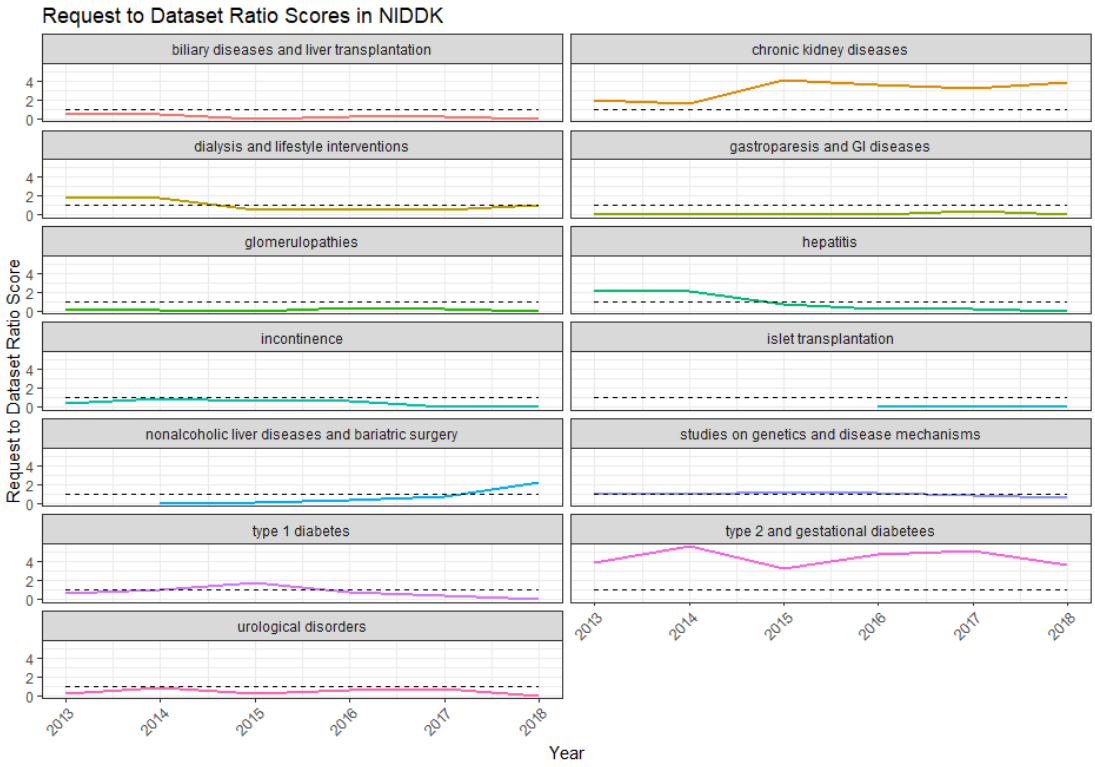

Figure 5-13. Request to dataset ratios for NIDDK datasets by topic, calculated annually from 2013 – 2018.

5.2.6 Summary of Findings

Considering requests of datasets by topic reveals that not all topics are requested at the same rate, with certain topics emerging as highly popular. Although some variation did exist over time, the RTD ratio for most topics stayed generally consistent across the range of years analyzed for each repository. Among all three repositories, the most highly requested topics were all related to illnesses and disorders with a significant disease burden (National Center for Health Statistics, 2017). For example, heart disease is the number one cause of death in the US, and had the highest ratio for topics in both dbGaP and NHLBI. Diabetes and chronic

kidney diseases, the two topics with the highest ratios in NIDDK are also both on the list of top ten causes of death in the US.

However, other topics are surprisingly under-requested based on their disease burden. For example, breast cancer is the most common cancer in the US, with more than 260,000 women diagnosed annually, and also the fourth deadliest, but falls close to the bottom of the rankings for the cancer-specific dbGaP requests (National Cancer Institute, 2019). Some of the multiple cancer datasets that are the most highly requested likely do contain some breast cancer data (for example, the highly-requested TCGA dataset does include breast cancer cases). However, datasets covering prostate cancer, which is also included in the TCGA dataset, receive requests at a rate 1.2 times higher than those covering breast cancer, despite breast cancer killing nearly 30% more people annually. The difference in requests might suggest that breast cancer is less studied in comparison to prostate cancer, but the opposite is true; in Fiscal Year 2017, NIH funded nearly 3 times more breast cancer research than prostate cancer research – $689 million for breast cancer and $239 million for prostate cancer (National Institutes of Health, 2018a), suggesting that breast cancer is in fact more widely researched than prostate cancer. Thus, it is not clear why prostate cancer datasets are requested so much more than those on breast cancer, which is both deadlier and more highly funded.

One possible explanation for the disparity between these cancers' dataset request rates and their disease burden and research funding could be that prostate cancer researchers receive less funding and therefore take advantage of existing

datasets to make the most of their limited funds. Conversely, prostate cancer researchers could be requesting less funding *because* of the fact that datasets that are suited to their purpose already exist and they therefore do not need to request funds to gather new data. The request data here does not provide enough information to draw a conclusion about the reasons behind this disparity, but does suggest potential avenues for future research.

Considered in combination with the results of the temporal analysis described in Section 5.2, these findings suggest that dataset requests do follow potentially predictable patterns and are not simply a function of datasets accruing requests over time. As will be discussed in Chapters 6 and 7, these findings have potential implications for how datasets are curated and stored.

5.3 Conclusions and Summary of Findings

This chapter has turned to information about the datasets themselves to better understand the dynamics and patterns of why some datasets are requested more than others. These analyses provide answers to the when and what of biomedical data reuse: when in a dataset's life cycle does reuse happen, and what are the topics that are most highly requested?

A number of factors appear to be at play in determining which datasets researchers choose to request. As would be expected, the age of a dataset has some influence in the number of total requests it has received – a dataset that has been around longer has had more time to accrue requests. However, at least among the

dbGaP datasets, age alone does not fully explain the number of requests a dataset receives. dbGaP datasets appeared to follow a cumulative advantage model, with the number of requests a dataset receives in its first year, regardless of when it was released, being highly predictive of how many requests it will receive in later years of its life. Some of the variation in request rates is also likely due to the topics that datasets cover. Among all three repositories, the most highly requested datasets were those that were related to common diseases that take a high toll on public health, with fewer requests for datasets covering rare diseases.

The findings presented here build upon the analyses described in Chapter 4 and help to provide a deeper understanding of how biomedical datasets are reused. Chapter 6 will discuss how the findings presented here and in Chapter 4 can be interpreted within the broad context of biomedical data reuse and explore what these findings tell us about who is using data and why.

Chapter 6: Discussion

The previous chapters have provided a view of the impacts of shared datasets – who is reusing them, for what topics they are being reused, when in their life cycle they are requested, where in the world they are being reused, and why they are reused. In this chapter, I will interpret the major findings of this study and discuss how these findings help advance our understanding of biomedical data reuse. I will also discuss the limitations of these findings and the context within which they can be meaningfully applied.

6.1 Summary of the Major Findings

This study aimed to provide a better understanding of how data are reused by exploring four broad research questions. Because research into data reuse is still nascent, I drew on an understanding of other phenomena in scientific research to formulate hypotheses for these questions. One exception to this is Research Question 4, about the topics of datasets that are most highly requested; little exists in the way of prior studies that would enable forming a hypothesis on this exploratory question. Table 6-1 provides a summary of the major findings.

Table 6-1. Summary of the major findings.

Research Question	Hypothesis	Finding
Research Question 1.1: For what methods and analysis types are datasets reused?	**Hypothesis 1.1:** Genomic datasets of the type found in dbGaP will be more likely to be used in combination in meta-analyses, while clinical	**Confirmed.** Genomic datasets from dbGaP are more often used together in meta-analysis and clinical datasets from NIDDK are more often

Research Question	Hypothesis	Finding
	datasets of the type found in the NIDDK repository will be more likely to be used on their own to answer an original research question.	used on their own for an original study. There are statistically significant differences in the ways that dbGaP and NIDDK datasets are used.
Research Question 1.2: How closely are the topics for data reuse aligned with the topics for which the data were originally collected?	**Hypothesis 1.2:** Similarity between original topics and topics of reuse will be lower for genomic data (found in dbGaP) than for clinical data (found in the NIDDK repository).	**Confirmed.** Similarity between original topics and topics of reuse is lower for genomic datasets from dbGaP than for clinical datasets from NIDDK. This difference is statistically significant.
Research Question 2.1: Where are requestors located in the world?	**Hypothesis 2.1:** Requestors will be primarily located in regions with a greater proportion of research institutions, including North America, Europe, and Asia.	**Partially confirmed.** Requestors are located around the world, but English-speaking countries are most over-represented when considering their requests compared to their international research presence.
Research Question 2.2: Are there patterns in career stage of requestors?	**Hypothesis 2.2:** A broad range of career stages, from student to full professor (or equivalent) will be represented.	**Partially confirmed.** While requests do come from a broad range of requestors, the majority of requests come from established researchers, rather than those early in their career.
Research Question 3: Are there temporal patterns to dataset requests?	**Hypothesis 3:** Patterns of requests relative to the original dataset release date will demonstrate a cumulative advantage process, similar to other scientific communication processes such as article citation.	**Confirmed.** Patterns of requests do appear to follow a cumulative advantage model, with patterns of requests over time similar to patterns of article citations over time. Early requests are predictive of later requests, especially for dbGaP.
Research Question 4: Are	NA	Datasets that contain data

Research Question	Hypothesis	Finding
there dataset topics that are more highly requested?		on more common diseases are more requested.

6.2 Interpretation of the Major Findings

This study used a variety of methods to describe biomedical data reuse to better understand patterns of reuse and the impacts of shared data for the biomedical research community. Throughout, I have framed this approach as providing answers to the questions of the who, what, when, where, and why of biomedical data reuse. Here, I interpret what the findings of this study can tell us about each of those questions.

6.2.1 Who is Reusing Data?

As I supposed in Hypothesis 2.2, researchers from across the research career life cycle reuse biomedical research data, from students just kicking off their careers, to mid-career professors, to well-established researchers and high-level commercial executives. This finding suggests that data sharing is more equitable in its current form – that is, in data repositories – than it had been through the interpersonal "gift economy" that previously characterized data sharing (Wallis et al., 2013; Yoon, 2017; Zimmerman, 2007). Students and early career researchers who would have lacked the professional network and status to be able to locate and negotiate access to data on their own can, and as this study found, do, make use of the datasets shared through repositories. These earlier career researchers can particularly benefit from the ability to use existing data, since they likely have less access to funding and other research

resources. The representation of researchers from both earlier and later career stages here suggests that a system of sharing data through repositories is more equitable and can help democratize research.

However, it is also notable that these various career stages are not *evenly* represented among requestors. Just under half of requestors to both the NIDDK and dbGaP datasets were established researchers at the full professor, senior scientist, executive, or director level, while assistant professors accounted for around a quarter of the requests. While these early-career and established researchers were making many requests, surprisingly, researchers in the middle of their careers were making fewer. Considering the relative difference in composition of requests for academic career stages (for which actual counts of researchers at each level are known) reveals that the number of researchers in a given career stage alone does not account for differences in rates of requests. The reason for the lower request rates among mid-career researchers cannot be determined with the data available in this study; further research could help elucidate the drivers behind different request rates.

Another finding that merits further exploration is the differences in rates of requests for associate and assistant professors to the dbGaP and NIDDK repositories. As discussed in section 4.2.2, associate professors are overrepresented in requests to dbGaP and underrepresented in requests to NIDDK, while the opposite is true for assistant professors. Further research would be needed to explain the reasons behind this finding, but it is possible that assistant and associate professors are engaged in substantively different types of research. The higher request rate to dbGaP for

associate professors could indicate that they are doing more genomic research or conducting more meta-analyses (the most common type of reuse for dbGaP data), while assistant professors' higher request rate to NIDDK could suggest that they are doing more clinical research or doing more original research studies (the most common type of reuse for NIDDK data). Analysis of the articles that arise from reuse among these two groups could help provide insight into how early versus mid-career researchers are reusing data.

6.2.2 What Are the Most Requested Topics?

The three repositories considered here include datasets covering a wide range of topics. Even within the NIDDK and NHLBI repositories, which are more constrained in terms of topic coverage than dbGaP, many different diseases and conditions are represented. In general, datasets about more common diseases and conditions were more requested than those that covered rare diseases. It stands to reason that a disease such as type 2 diabetes, which affects may people, would be the focus of more research, and therefore receive more reuse requests, than something such as a rare genetic disorder that affects only a few families in the world. On the other hand, the datasets that cover uncommon diseases do not go entirely unrequested, suggesting that they still represent a valuable source of data for the researchers who are engaged in such research.

Disease burden and research density alone do not fully explain request rates for some topics. The example discussed in section 5.2.5 – the relative request rates of

prostate cancer and breast cancer datasets – demonstrates that not all topics are requested at a rate that correlates with the relative disease burden and research funding for that topic. Based on this analysis, which simply compares relative request rates, it is impossible to know what other factors might be at play in determining what topics researchers are most likely to request. Perhaps prostate cancer data are more difficult or expensive to collect than breast cancer data, and therefore researchers are more likely to request existing datasets rather than collect their own. Perhaps the prostate cancer datasets have been cited more in the literature, thus giving them higher visibility. Perhaps the prostate cancer datasets just happen to be better described and more clearly documented than the breast cancer datasets and are therefore more useful. Perhaps it is an issue of gender disparity in research, with prostate cancer, a disease affecting men, receiving more requests than breast cancer, a disease primarily affecting women. These findings suggest that further research into the broader funding, publication, and disease context in which datasets are requested could provide additional insight into the drivers behind the patterns of requests by topics seen here.

6.2.3 When in a Dataset's Life Cycle Are Requests Made?

Temporal analysis of data requests reveals that long-term requests of datasets can likely be predicted from early requests. In both the dbGaP and NHLBI repositories, the number of requests that a dataset receives in the first three years after its release is a good predictor of how many requests it will receive in the long-term,

considering both total requests and requests made after the first three years. This finding holds true even when controlling for age, suggesting that the number of requests a dataset receives is not merely a function of how old it is. However, interestingly, while the first year of requests is a good predictor of long-term reuse in dbGaP, it is actually a very *poor* predictor of reuse in NHLBI. It is not until the second year that NHLBI datasets begin to be requested at a rate that is predictive of long-term reuse. This finding could be due to differences in patterns of how clinical versus genomic datasets are reused, or could be reflective of differences in how datasets from these two particular repositories are reused. Unfortunately, the NIDDK repository did not have enough historical data to include in this analysis, which would have provided a means of better understanding whether the difference could be ascribed to differences in ways clinical data is used. However, this analysis could be expanded to include other repositories to determine the mechanism behind these different request patterns.

Within dbGaP, datasets also follow typical patterns of request over the course of their life cycle that suggests that dbGaP data reuse requests, much like other scientific processes, follow a cumulative advantage model – success breeds success. Datasets that are highly requested early in their life go on to continue to be highly requested, whereas datasets that receive few requests in their first years tend to continue to be less requested. Dataset requests are also similar to article citations in that they tend to reach a peak number of requests and then receive gradually fewer requests over time. In article citations, this peak is often achieved around five to ten

years after the article's publication (Wang, 2013); for datasets, this peak occurs in the second year of the dataset's life, after which requests slowly taper off over time.

The shorter time period in which datasets reach their peak compared to articles could be due to differences in where in the life cycle of datasets requests happen versus where in the life cycle of articles citations happen. A use request happens much earlier than an article citation, at the start of the research process rather than at its end. The publication process often stretches over the course of months, as the article goes through peer review, potential revisions, and preparation of the final documents, so there will always be a lag between the time that a researcher uses an article and the appearance of evidence of that use, in the form of a citation. On the other hand, a use request provides evidence of use immediately. Therefore, it is likely that patterns of when datasets and articles are *used* are similar, and what differs is just the times at which evidence of that use appears.

A surprising exception to the finding about peak request year was that the most highly requested datasets – those in the 90th percentile of overall requests – diverge from the pattern observed in the less-requested datasets in ways that suggest different dynamics could be at work in driving requests. The mean number of requests for these datasets does reach a peak and then drop off in the third year after release, like datasets in the other percentiles. However, the mean number of requests begins to rise again in the fourth year, increasing over subsequent years and eventually even surpassing its previous peak. Without further research, it is difficult to definitively say why this pattern occurs, but one possible explanation is that the

most highly requested datasets see peaks at the usual time for datasets and the usual time for article citations. That is, the dataset is released, and, like its peers, reaches its peak in the first year, following whatever dynamics drive requests. However, unlike less-requested datasets, these highly requested datasets *also* go on to be cited in articles reporting on the secondary reuse that arose in this first wave of requests. As descriptions of dataset reuse start to appear in the literature, perhaps the temporal pattern of requests starts to behave more like article citations, reaching a peak around the same time that the article describing the dataset would also expect to see a peak in citations (around 5-10 years after publication of the article).

With only ten years of requests available for this analysis, it is impossible to know whether this explanation holds. Mean requests for 90th percentile datasets were at their highest in the final year of requests available for this study; without additional years of data to consider, it cannot be known whether that year is the peak or whether requests will continue to increase over time. Revisiting this analysis a few years from now, when additional years of requests are available, could demonstrate whether in fact this predicted pattern occurs. In addition, having a better mechanism to connect datasets with articles that cite them would help provide additional evidence that could support this potential explanation. At present, data citation mechanisms do not allow for sufficiently accurate counts of articles that cite datasets to enable a meaningful analysis of this theory.

If further research supports the initial findings of this study – that the number of requests a dataset receives early in its life is predictive of its long term reuse –

reward and credit for researchers who share highly reused datasets could come more quickly than with other measures of scientific success or productivity. One criticism of measures such as citations to articles is that these are lagging measures that cannot show impact until months or even years after the release of the original article. As discussed above, the nature of the scholarly publication cycle means that citations to articles generally do not even begin to appear until well after the article's publication, peaking sometimes as late as a decade after the article's publication. Some bibliometricians have tried to identify measures that could provide earlier identification of high-impact articles, so-called altmetrics such as mentions of the article on Twitter or number of times readers have saved the article in Mendeley. While altmetrics provide quantitative counts of attention early in an article's life cycle, that attention generally does not translate into long-term impact in the form of article citations, limiting their usefulness as a means of assigning meaningful scholarly credit (Thelwall et al., 2013). If it can be demonstrated that early attention to datasets in the form of requests in the first few years *do* reliably predict long-term use of datasets, credit could be given comparatively early in the research life cycle to researchers who share high-value datasets. Being able to recognize researchers who share high-value datasets soon after they share them, rather than having to wait years to receive credit, could incentivize researchers to not only share datasets, but to do so in a timely fashion.

6.2.4 Where in the World Are Requestors Located?

Although the repositories considered here are funded and administered by various organizations within the NIH, an agency of the US government, the datasets contained within them are available worldwide and represent a potentially valuable global research resource. Indeed, requests do come into these repositories from all around the world, but the global distribution of requests is far from uniform. Even when accounting for research presence by considering the number of universities within countries, the United States is highly overrepresented in requests to all three repositories.

Outside of the United States, other patterns in which countries were over- and underrepresented emerged. Other English-speaking countries such as Canada, the United Kingdom, Australia, and New Zealand, were also overrepresented given their share of universities. This is a finding that I did not predict, but it is logical given that that the websites and documentation for all three repositories considered here are available in English only, making it more challenging for non-English speakers to request and meaningfully use the data. Applying a similar methodology to analyze geographic distribution of reuse for datasets in repositories documented in other languages could provide a comparison to test whether researchers' native language drives their choice to reuse certain datasets.

Besides potential language barriers, other geographic factors may influence rates of reuse among researchers in various countries. Researchers may be more familiar with repositories located within their home country or region than those in

other parts of the world. Previous studies on researchers' data reuse practices have identified trust in the repository as a major factor in the decision to reuse data (Faniel & Jacobsen, 2010; Faniel et al., 2015; Rolland & Lee, 2013; Yakel et al., 2013; Yoon, 2014, 2017); perhaps researchers are more likely to trust a repository located within their region. One example that supports this hypothesis is the existence of an international collaboration among three nucleotide sequence databases: GenBank, located in the United States; the European Molecular Biology Laboratory (EMBL), with several locations in Europe; and the DNA Data Bank of Japan (DDBJ) all contain the exact same data. The three databases are synchronized daily, so that a user need only submit data to one of the databases for it to be available in all three (Baker et al., 2000). While distributed and redundant data storage makes sense from a preservation perspective, the fact that these three identical databases exist with their own distinctive names (two of which reference geography explicitly) suggests that researchers might make choices about where to look for data based on the geographic location of the repository. These three repositories do not require submission of a use request to access data, so other methods would be needed to track patterns of reuse, such as analysis of use logs and IP address access, but such an analysis could provide insight into the extent to which geographic factors play a role in researchers' choice to use data from a repository.

For all three repositories in this study, North American and European countries (including European countries where English is not the official language) were the most overrepresented, while countries in Asia, the Middle East, Africa, and

South America were almost universally underrepresented – if they were represented at all. The low use of data in Asian nations in particular was a surprising finding, given the major research presence within that region. For example, together, India, China, and Indonesia have about 30% of the world's universities, yet account for only 2% of the data requests. The majority of countries in the world that had at least one university had no requests at all to any of the three repositories. This finding suggests that these valuable data resources might not be benefitting the researchers who could potentially gain the most value from them: those in countries with less research funding and therefore less resources with which to collect their own data.

Within the United States, requests are more evenly distributed among states than they are among countries in the world. There were fewer extremes among requests within the US, with most states requesting data from the three repositories at rates that are in accordance with the amount of NIH funding received by research institutions within the state. The less extreme variations between request rates and research presence within the states versus within countries could simply indicate that the proxy for research presence within states – NIH funding – is a better representation of biomedical research presence than the proxy used for countries – number of universities.

However, outliers do still exist – Alaska, New Mexico, and Wyoming are somewhat surprising outliers in terms of overrepresentation. This finding could indicate that these states are requesting more data than might be expected given the amount of research underway, but conversely, it could also mean that these states are

receiving less NIH funding than might be expected given that amount of research. Perhaps researchers in these states are unable to secure adequate NIH funding to support large-scale data collection, so they turn to existing datasets to fill the gap. On the other hand, it is possible that these researchers simply are not applying for as much research funding because they are already planning to reuse existing data. Analysis of not just the research that is funded in each state, but what proposals are *not* funded could help elucidate the reasons behind the funding/request discrepancy (although information about unfunded proposals is not publicly available). Either way, this finding could support economic arguments in favor of sharing and reuse of biomedical research data – not only can reuse of data save money that would have otherwise been spent on gathering new data, but sharing also increases the return on investment of scientific research funding by extracting additional discovers from the original data (Arzberger et al., 2004; Costello, 2009).

6.2.5 Why Are Requestors Reusing Datasets?

This study revealed that researchers request data for a variety of different reasons – sometimes they simply want a dataset in which to test a research question, but researchers also request data to pool multiple datasets for questions that one dataset alone cannot answer, develop and test new statistical methods, design and validate software and computational tools, develop data infrastructure, and more. While any given dataset can and often is used in a variety of different contexts, the genomic and clinical datasets here demonstrate different patterns of reuse that are at

least in part accounted for by the different methodological limitations and practices associated with these two data types.

As section 4.1.1 describes, some of the methodological differences in how researchers use datasets can be explained in part by the strengths and limitations of certain types of data. Genomic dataset of the kind found in dbGaP often must be combined with each other to achieve the massive sample sizes that are needed to achieve adequate statistical power for this type of study (Hong & Park, 2012). With this in mind, the genomic research community has developed data standards to ensure that, wherever you are in the world and whatever type of equipment you use to collect your data, it will likely be interoperable with other genomic datasets (Field et al., 2011).

On the other hand, clinical datasets of the kind found in NIDDK often use variables developed uniquely for a specific research study, often aimed at capturing subjective measures of patient experience. Similar concepts may be represented with varying degrees of difference among studies, such as the example discussed in Chapter 4 of differences in how alcohol consumption and binge drinking are defined in two similar studies from the NIDDK repository. Even if the discrepancy between how two studies define a concept is slight, those two datasets cannot be meaningfully combined. Although the NIH has made efforts to encourage the use of Common Data Elements (CDEs) that would enable harmonization of data across studies, uptake has not been universal, and researchers will still face problems integrating datasets that

have already been collected without CDEs, such as many of the datasets in NIDDK (Sheehan et al., 2016).

These differences in how data are used also influence the degree of similarity between the topic of the original dataset and the topic for which it will be reused. NIDDK datasets were used in contexts that were similar to the original reason for which the data were collected. Over half of request/dataset pairs had a semantic similarity score of 1, meaning that the request proposed reuse in the exact context for which the data had been originally collected, and the mean score for NIDDK was 0.78, demonstrating a high degree of similarity between reuse and the original data context. This finding makes sense, considering the attributes of clinical datasets described above. These datasets focus not only on a defined patient population, but also on fairly specific characteristics of that population – their response to a particular drug or intervention, symptoms and clinical findings related to their disease, or their self-described perception of their health and emotional well-being. While these datasets provide a depth of understanding – often featuring hundreds, if not thousands, of variables – they provide it in a very specific context, meaning that the applicability of these datasets is relatively constrained to a small set of related topics.

On the other hand, genomic data is comparatively uncomplicated, consisting of the genetic sequences of individuals with a certain condition (or even normal, healthy individuals). Not only are these data interoperable with other genomic datasets, but they are also more generally applicable beyond a narrow disease category. As a result, they are used in a broader range of reuses that may diverge

quite significantly from the original reason for which they were collected. The mean semantic similarity score for dbGaP request/dataset pairs was only 0.56, and nearly a third of them had a score of 0, meaning that the request proposed a topic of reuse that was completely different from the reason for which the data had been collected.

It may be tempting to suggest that dbGaP datasets are more useful than NIDDK datasets, since they are not only more requested, but also reused in a broader range of contexts. As Chapter 7 will discuss, just because a dataset is infrequently requested does not mean that it lacks value. However, the datasets that are most likely to be requested frequently and for the broadest range of reuse may merit additional curation or prioritization for preservation.

6.3 Methodological Contributions of the Study

This study is an early exploration of questions that need to be answered to understand the impact of data sharing and thereby reward researchers who share high-value datasets. As this study has demonstrated, data reuse takes many forms, and also introduces a set of methods for understanding various aspects of this complex phenomenon. These methods will also be of use to repositories who wish to better understand who is using their data and how. Researchers could also benefit from knowing these answers to these questions as well, so repositories could consider creating dashboards or reports that draw on these methods to provide more detailed information to researchers beyond simple counts of reuse.

First, this study introduces semantic similarity as a method to understand how similar a proposed reuse is to the reason for which the dataset was originally collected. Using MeSH terms is a useful approach here, since the datasets already have MeSH terms, and the availability of a reliable automated text indexer, which NLM makes freely available, enables easy description of texts with little manual intervention. Because semantic similarity is used in a range of biomedical text comparison applications, packages exist for R and other popular statistical software, incorporating existing, validated algorithms, lessening the challenges of adoption of semantic similarity as a metric. While measuring semantic similarity with MeSH terms is limited to texts within the context of biomedical literature, other similar methods exist for quantitatively determining similarity between a pair of texts, so repositories with other types of data could use either a discipline-specific or a general-purpose measure.

The coding of reuse requests in this study gives new insight into the ways that datasets are reused by expanding on the existing taxonomy of reuse types drawn from the literature. This expanded taxonomy provides a more complete understanding of the ways that datasets are reused and is validated by external coders. While other types of reuse likely exist outside of biomedical research, this taxonomy provides a basis for categorizing and understanding types of reuse. Unfortunately, this method is time-consuming because it requires manual coding of reuse requests, which can only be done by someone with a reasonably comprehensive understanding of the science described in the requests. However, in future research, I intend to use the set of use

requests I have coded with reuse type as a corpus for a machine learning text classifier to determine whether an automated approach could be used to categorize requests, which could replace the manual process, at least in the context of repositories with similar types of data to those discussed here.

This study draws from a discipline quite distant from biomedical research, borrowing the measure of relative difference in composition that is used to assess racial and ethnic disproportionality in educational settings. This metric moves beyond raw counts of reuse to contextualize the extent to which researchers from particular countries or career stages are reusing existing datasets. I have used number of universities per country and amount of NIH funding by state as a proxy for research presence, but this method could also be used with other ways of approximating research presence, such as funding from another relevant funder or number of publications arising from a country.

To understand temporal patterns of datasets over time, this study proposes two techniques: tracking use by deciles of overall reuse and quintiles based on the mean decile per year over the course of the dataset's life. This method can be applied to dataset requests from any repository, regardless of discipline, since it does not rely on information about the dataset or topic of request. Further, this method could also be used to explore cumulative advantage processes outside of dataset requests, such as citations to articles over time.

Finally, this study introduces the request to dataset ratio as a way of understanding which topics are most requested. This method could also be applied to

other repositories in different disciplines or even to other comparisons of topics, such as comparing citations to articles with certain topics. Here, I use a topic modeling algorithm to identify topics within the datasets, a technique that is broadly applicable to texts of any type, regardless of their linguistic content. This approach could therefore be applied to any repository, but topics could also be determined manually or by drawing on metadata from the dataset descriptions. For example, because the topic model did not perform well for the dbGaP datasets, I used primary phenotype to determine the topics. Once topics are determined, the request to dataset ratio can be used with any number of topics and any number of datasets to provide insight into the topics that are most requested.

Based on the variation in findings among the three repositories studied here, study of repositories from other disciplines would also likely exhibit some differences in how datasets are reused. In addition to providing a set of methods for exploring data reuse, this study provides a set of data to compare against to understand how reuse differs from one discipline to another, or even from one repository to another within the same discipline. This study also provides a baseline against which to compare data reuse over time. For example, it could be informative to revisit these analyses after the NIH implements its forthcoming data management and sharing plan policy to determine whether the increased demand to share datasets impacts reuse.

6.4 Limitations and Considerations for Application of Findings

As has been discussed, this study aimed to provide a preliminary understanding of a very complex phenomenon. As such, the findings should be understood and interpreted in that context. This study considers a very small group of repositories, several of which had incomplete data (such as NIDDK, which was missing release dates and request info from before September 2013, or NHLBI, which did not provide me with data on requestors or the text of use requests). Even where full data were available, the NIDDK and NHLBI datasets had much fewer requests than dbGaP, so these findings should be considered with less certainty, given that any variations here may be due more to the smaller population size than to actual differences in the phenomena described.

As has been discussed, reuse of data is difficult to quantify. This study used requests to reuse data as a proxy for reuse, which is likely a better proxy than some other measures, such as download counts or citations within the scholarly literature, but they are still an imperfect measure. Although requestors must have a fairly specific reason for which they intend to use the data, their actual research may not proceed according to those plans. A researcher might request a dataset and then, upon receiving it, discover it is not actually suited to her needs after all and end up not using it. Anecdotally, researchers have told me that the request process for some of the repositories is onerous enough that they sometimes request more datasets than they will likely need just in case, rather than find out later that they need additional data and have to go through the process again. Connecting use to requestors may also

lead to inaccuracies in understanding the career status of reusers; the person who requests the data may not actually be the person reusing it. A professor might request a dataset on behalf of a student, or a project manager on behalf of a research team. Therefore, a data request cannot be considered exactly equivalent to an instance of data reuse, and results should be interpreted with this consideration in mind.

As this study has demonstrated, findings that hold true for one repository may not hold true for another, which suggests that the ability to generalize findings across repositories may be limited. Some of the findings were similar across repositories – datasets were almost universally most highly requested by researchers in the United States and other English-speaking countries, and topics with significant global disease burden were among the most requested for all three repositories, compared to rare diseases. However, for other questions, the findings differed widely between repositories. For example, the types of research for which dbGaP and NIDDK datasets were used differed widely, as did the temporal patterns of use between dbGaP and NHLBI. That this much difference existed between three relatively similar repositories – all three housing human subject data related to biomedical research and funded by the NIH – suggests that data reuse is not a phenomenon with simple, universal explanations. Therefore, caution should be used in trying to apply these findings to biomedical research repositories or datasets more broadly, and they almost certainly should not be applied to data and repositories from other disciplines.

6.5 Summary of Discussion

This chapter has provided an interpretation of the findings, with a particular focus on what this study can tell us about the who, what, when, where, and why of data reuse. The answers to these questions help extend our understanding of the nature of biomedical data reuse and contribute to the development of scholarship in this area. This study was designed around a specific definition of reuse and constrained by the limited information that is currently collected about data reuse, so these findings must be interpreted within the context of a specific type of biomedical data reuse. Despite these limitations, these findings suggest potential implications for a range of stakeholders in the biomedical research ecosystem, which will be discussed in Chapter 7.

Chapter 7: Conclusion

With researchers increasingly being required to share their data, the amount of publicly available and potentially reusable biomedical research data will continue to grow. Understanding how those datasets are reused will help ensure that informed decisions are made about how to best curate, preserve, and share data, as well as how to reward researchers who share high-value datasets. Shared datasets exist within a complex research ecosystem with a variety of stakeholders; accordingly, I will suggest how each of these stakeholders could consider acting on the findings of this study. Given the exploratory nature of this study, I will also propose future research that could build upon, confirm, and explain the findings I have presented within this dissertation.

7.1 Implications of the Findings

7.1.1 For Researchers

The findings of this study may help allay some of the concerns that researchers have expressed about sharing their data. Researchers have worried that they might get "scooped" if they share their data – that someone else will beat them to publication on a discovery that they would have gone on to make themselves (Laine, 2017). One controversial editorial on data sharing worried that researchers who reuse data would end up "possibly stealing from the research productivity planned by the data gatherers" (Longo & Drazen, 2016, para. 3). However, the findings of this study

suggest that the ways in which researchers are reusing shared data make it unlikely they will end up scooping the original data collector in most cases. Especially for data in dbGaP, the context in which researchers proposed to reuse datasets often diverged markedly from the reason they were originally collected. These reusers are unlikely to scoop the original data collectors because they are looking at such different questions than the collectors were.

While topics of reuse were more similar in the NIDDK repository, only about half of the request/dataset pairs had a semantic similarity score of 1, meaning they were reusing the data in the same context as the original collector. Of course, a semantic similarity of 1 does not mean that the reuser is doing the exact same research as the original collector. Semantic similarity scores are based on the MeSH terms assigned to use requests and datasets, which are mostly diseases or even broad disease categories. A use request and its corresponding dataset would have a semantic similarity score of 1 if they were both described as covering "Kidney Diseases," but this term is sufficiently broad that the reuse and the original study could actually be considering quite different questions. Even so, clinical data does generally have more limited reuse potential than genomic data, based on the type of information contained in these datasets and how it is collected. The potential to be scooped is therefore perhaps higher for researchers sharing clinical data than those sharing genomic data.

It should also be noted that not sharing data does not protect a researcher from being scooped; it happens all the time and did even before sharing data became a common practice. The nature of scientific research and discovery means that there

are often multiple research teams around the world working on a topic at any given time, not because one is riding the others' coattails, but simply because "we tend to make important new advances when the tools (intellectual and technical) become available, and others are not unlikely to do the same" (Mole, 2004, para. 10). In fact, in some cases, data sharing and other open science practices can actually help *prevent* scooping by establishing the primacy of one's scientific claim. For example, researchers may choose to pre-register their studies using a platform such as the Open Science Framework (where data can also be shared), a process by which they state in advance their outcomes of interest. Because pre-registering or sharing data in a repository creates a time stamp, researchers can definitively demonstrate that they were the originator of an idea or discovery, helping to lessen the possibility that they will be scooped (or at least giving them ammunition to fight back if they are).

Other researchers have expressed concern that making their data publicly available might open them up to scrutiny of their original results by outside researchers (The International Consortium of Investigators for Fairness in Trial Data Sharing, 2016). With increasing concerns about the reproducibility of research, this concern is not entirely unfounded ("Reality check on reproducibility," 2016), although one might argue that making sure your original results are correct before publishing might be the best course of action to avoid such problems. While it may seem that re-running analyses on the exact same dataset would necessarily lead to the same results and outcomes, it often turns out that this is not the case; it is entirely possible to use the exact same data and come to entirely different results, particularly

when the original authors have not clearly documented the computational methods they have used in their analysis. Results can be dependent on factors such as the specifics of the computing environment, software versions and dependencies, and choices the researcher makes about parameters of the analysis (Begley & Ioannidis, 2014; Grüning et al., 2018; Stodden et al., 2012).

This study's results suggest that, at least for dbGaP and the NIDDK repository, reproducibility studies or other attempts to replicate the original study's findings are not common purposes for requesting the data. Only 11 requests for dbGaP and two for NIDDK indicated they intended to use the data to reproduce the original results. These numbers correspond to just 0.05% of all requests for dbGaP data and 0.36% for NIDDK data. Of these requests, most described an interest in reproducing the results using slightly different methods, such as using different software or different sampling criteria, rather than questioning the original findings. Only one of the requests indicated that it aimed to re-analyze the data because the original findings had not been confirmed in other studies; the requestor speculates that this finding "was a spurious result of inappropriate statistical technique." However, this request is only one out of thousands, indicating that reanalyzing data for the purpose of debunking the original findings is not a major type of reuse.

Of course, this is not to say that reproducibility is not a significant problem in biomedical research; many researchers have raised alarms over the reproducibility of biomedical research (Begley & Ioannidis, 2014; Ioannidis, 2005, 2014). A range of efforts are underway to increase reproducibility in biomedical research, such as

development of guidelines to enhance (National Institutes of Health, 2017b) and tools to encourage broader adoption of open scientific practices (Munafò et al., 2017; Nosek et al., 2015a; Nosek & Bar-Anan, 2012). However, it appears from this study that verifying or reproducing results is not a common use of shared research data. Limiting access to data based on an individual's concern about possible scrutiny when sharing has the potential to further science and enhance human health does not serve the public good, particularly given that the findings of this study suggest that this type of reuse is rare.

7.1.2 For Repositories and Curators

Patterns of use requests – both temporal patterns and patterns of highly requested topics – can provide an evidence base for informing curation and preservation decisions. While it may seem desirable to preserve *all* biomedical data indefinitely, just in case it is of use at some point, doing so is not feasible, nor would long-term storage of certain datasets be an efficient use of funds. For example, as costs of genome sequencing continue to decline, in some cases it may actually be cheaper to just re-collect data rather than store them (Weymann et al., 2017). Curating data to ensure they are in a usable and discoverable form often requires significant human effort, and despite decreasing costs of memory and the availability of cloud storage, long-term preservation can come with high costs. The findings of this study are preliminary and do not hold across all three repositories, but at least for the data in dbGaP, the number of requests a dataset receives in its first year is highly predictive

of the number of requests it will receive over the long term. It may therefore be possible to make meaningful curation decisions early in the data life cycle, prioritizing the datasets that are most highly requested in their first few years.

In addition to predicting future use based on early request rates, it may also be possible to anticipate demand for datasets based on the topics they cover. As this study demonstrated, datasets that focus on common diseases are more requested than those that focus on rare diseases. However, that is not to say that datasets covering rare diseases should be discarded or ignored; in fact, quite the opposite is true. Even though they may be less requested than datasets on more common and well-studied disorders, data on rare diseases are in a sense more valuable because they are more difficult to re-collect. Given the prevalence of diseases such as heart disease, type 2 diabetes, and cancer, finding participants for studies on these topics would be relatively easy, since they affect so many people. On the other hand, it is much more difficult to locate patients with rare diseases by virtue of the fact that they are rare. Especially in the case of genomic research, which requires larger sample sizes, it is often necessary to pool rare disease data from multiple sites that are able to collect the data from small patient groups to whom they have access. Repositories have been described as "unequivocally essential" to rare disease research, given their important role in facilitating access to rare disease data that can support research that might not be accomplished otherwise (Raza & Hall, 2017, p. 37). In fact, bioethicists have argued that researchers have a responsibility to their participants to share research data, particularly in the case of rare diseases. These patients have freely given their

time and data to participate in research that they hope will lead to treatments, and researchers should do all they can to advance that work, including sharing data (Hansson et al., 2016).

Repositories must find a balance between focusing curation and preservation efforts on datasets with high reuse potential and those that may not be reused as often, but have value because of their rarity. Library practices may provide insight into how to prioritize curation and preservation of certain content without entirely discarding lower-use materials. For example, NLM provides enhanced indexing of certain journals that are searchable within its PubMed bibliographic database. The subset of journals that are selected for inclusion in MEDLINE (one of the underlying data sources searched within PubMed), based on criteria such as journal scope and coverage and quality of content, are indexed with additional metadata such as Medical Subject Heading (MeSH) terms and publication type (National Library of Medicine, 2019). Articles from journals that are not selected for MEDLINE indexing can still be searched in PubMed based on metadata such as keywords in their abstract, or author's name; they just do not have the added information that comes from the NLM's investment of a curator's time that enhances the metadata associated with selected journals.

Library practices can also provide guidance on how repositories might choose to make preservation decisions. Libraries must make choices about their collections based on the physical limitations of their space; there are only so many books that can fit on the shelves. Sometimes this means discarding items that are out of date,

damaged, or no longer used. This choice may be appropriate for some datasets in repositories as well, especially if technologies advance in ways that make existing datasets technologically obsolete. On the other hand, sometimes libraries have books that are not highly used, but still merit keeping, perhaps because they have historical value, or are still used from time to time. Off-site storage can provide a location to more cheaply and efficiently store less-used items, with a tradeoff in terms of convenience – a user must request the item and wait for it to be retrieved, rather than walking in to the library and simply taking it off the shelf. Repositories could take a similar approach of using "cold storage" for infrequently used data (Dell EMC, 2019). Cold storage methods are more economical and computationally efficient, preserving high-cost and high-performance systems for frequently accessed data while still enabling preservation of lower-use data. Researchers who want to use a lower-use dataset may have to wait a little longer to get it, but they will still be able to get access, while the repository can help control storage costs.

This study also demonstrated that biomedical data reuse is not evenly distributed among researchers around the world. Repositories could consider outreach to under-resourced regions to increase awareness of and access to freely available data resources. In many parts of the world, potential partners are already in place who could facilitate this outreach. For example, the NIH and other US funders support a variety of research and capacity building efforts in Sub-Saharan Africa (National Institutes of Health Fogarty International Center, 2019). Libraries and institutions that train researchers would be natural partners to help increase awareness and access.

Initiatives such as the Hinari Access to Research for Health Programme and Librarians without Borders, which already provide training and resources for librarians in underserved regions, could help to increase librarians' knowledge of how to support researchers interested in working with existing research data (Medical Library Association, 2019; World Health Organization, 2019).

Establishing contacts within those regions could also help encourage researchers to in turn deposit their data in these repositories, which could significantly increase the usefulness of the repository as a research resource. For example, the Human Health and Heredity in Africa (H3Africa) project aims to increase research infrastructure and expertise to collect genomic and clinical data from African populations (Human Health and Heredity in Africa, 2019). Repositories could greatly improve their representation by ingesting this type of dataset. A 2016 study found that over 80% of the existing genomic data in the world came from people of European descent; other populations made up as little as 0.05% of the existing genomic data (Popejoy & Fullerton, 2016). Partnering with researchers in other regions of the world could therefore not only increase access and use of existing data, but potentially create pathways to increase the diversity of subjects represented in repositories and thereby improve healthcare for patients of all races.

7.1.3 For Research Funders

As this study has demonstrated, biomedical data repositories represent a rich source of data to fuel research across a broad range of topics, sometimes diverging

widely from the original purpose for which the data were collected. The NIH has, accordingly, made a significant investment in curating and making available data arising from NIH-funded research. The recent NIH Strategic Plan for Data Science highlights the need to develop infrastructure and policies that help make biomedical research data FAIR (findable, accessible, interoperable, and reusable), thereby enhancing the ability of researchers to locate and reuse the data (National Institutes of Health, 2018b). The findings of this study suggest that researchers do have an interest in using shared data from repositories, and further emphasis by NIH on funding and policy towards increasing FAIRness of data could help increase reuse, as well as making reuse of data easier and lowering the barrier to entry for reusing data.

In addition to providing funding and policy guidance that will increase the availability and usability of biomedical research data, the NIH has also begun to encourage researchers to reuse data by providing funding specifically for that purpose. While many Funding Opportunity Announcements (FOAs) mention that secondary analysis or data reuse are permitted, a few of the currently active FOAs are intended specifically for that purpose. Some of these FOAs are specific to particular disorders or areas of research, such as "Secondary Analyses of Existing Alcohol Research Data" and "Cancer-Related Behavioral Research through Integrating Existing Data," or even fund use of data from a specific repository, such as "Leveraging Population-based Cancer Registry Data to Study Health Disparities," which funds secondary analysis of data in either the Surveillance, Epidemiology, and End Results (SEER) Program or the National Program of Cancer Registries (NPCR)

(National Institutes of Health, 2016, 2017a, 2018d). These FOAs highlight some of the benefits of reusing data – accelerating discovery, increasing cost-efficiency, and enabling access to large datasets or data on rare diseases that researchers likely would not be able to gather on their own.

While these FOAs can help raise awareness of existing data resources and incentive their reuse, it is important to caution that support for reuse of shared data should not be considered an alternative to providing funding for researchers that aim to collect their own data. For example, an NIH pediatric cancer research effort proposed in 2019 features data sharing as a major focus of the initiative. While cancer researchers generally recognize the importance of sharing and combining data, especially in the context of rare cancers, some argue that making data sharing the emphasis in this initiative is ineffective. They point to differences in the biology of childhood cancers that make integrating data from multiple sources a less meaningful approach than in the context of adult cancers and suggest that funding other approaches might be more effective than prioritizing data sharing (Kaiser, 2019). Not all questions can be answered with existing data, and as technologies progress, older data may no longer be useful. Therefore, even as more data of higher quality become widely available, reuse of existing datasets should be considered complementary to rather than a replacement for research activities that involve collecting new data. While this may seem so obvious that it hardly seems worth noting, I believe it bears explicitly stating given a political climate in which some government entities are seeking to cut funding to federal agencies that conduct and fund research.

Funders also have an important role to play in thinking about how they will not only encourage reuse of shared data, but also reward the researchers who originally collected datasets that go on to be reused. As Chapter 2 discussed, the notion of credit and reward are foundational to scientific norms (Carpenter et al., 2014; Durieux & Gevenois, 2010; Garfield, 2002; Holden et al., 1994; Kochen, 1987; Latour & Woolgar, 1986; Merton, 1942). Many researchers already balk at the idea of sharing data because they see it as giving away one of the products of their financial and intellectual investment (Longo & Drazen, 2016; The International Consortium of Investigators for Fairness in Trial Data Sharing, 2016). Funders (as well as research institutions) are in a position to encourage and incentivize data sharing by giving credit to researchers who have shared data that goes on to be used by others.

As this study has demonstrated, not all biomedical datasets are reused equally. Some of the datasets in this study had been requested hundreds or even thousands of times, whereas others only had a handful of requests. Part of the variability in the number of requests a dataset receives is due to the type of data it contains; for example, NIDDK clinical datasets, with their relatively constrained uses based on the way the data are collected, are requested less than dbGaP genomic datasets, which are more interoperable and tended to be used in a range of topics that diverged more significantly from the original context in which the data were collected. Given the differences in these types of data, it would be reasonable to expect that dbGaP datasets, which have a wider range of uses, would be more requested. Based on these differences, it hardly seems fair to compare datasets from these repositories based on

counts of requests alone; dbGaP datasets had 103 requests on average compared to just 8 requests on average for NIDDK datasets. Raw counts alone simply cannot be used to compare dataset use across multiple repositories.

Even within repositories, using raw request counts may be an ineffective means of rewarding data sharing, since dataset *use* may not always be equivalent to dataset *value*. As the topic request analysis in section 5.3 demonstrated, datasets that cover common illnesses receive more requests than datasets covering rare illnesses. However, it could be argued that a dataset on a rare disease is more valuable than one on a common disease, regardless of how much either dataset is used. As discussed above, data on rare diseases is more difficult to come by and would be more difficult to recreate than data on common diseases, which have plenty of potential subjects to draw on. It could be reasonably argued that a researcher who shares a dataset on a very rare disease is making a significant contribution to research and to meaningfully improving the lives of patients who would not otherwise have been the focus of research beyond of the original researcher's work, even if only a few other researchers use the data. To suggest that such a dataset deserves less credit than a dataset that is requested many times risks rewarding researchers of common diseases over researchers of rare diseases and could even potentially disincentivize sharing of rare disease data.

Much as article citations are a flawed means of measuring the actual value or impact of an article (Edwards & Roy, 2017; Lane, 2010; Werner, 2015), simple counts of dataset reuse (whether that is measured by requests or other quantitative

counts) is likely an inaccurate means of determining a dataset's impact. Bibliometricians have begun to call for responsible application of metrics to avoid creating perverse incentives or misunderstanding the actual impact of articles, and that field has long been characterized by efforts to develop more accurate means of measuring and quantifying scientific impact (Edwards & Roy, 2017; Hicks et al., 2015). The scientific community has a rare opportunity now, as data sharing begins to become a more standard and formalized practice, to think carefully about how data sharing should be quantified, considering such questions as how value in data is defined and how to give credit for sharing in ways that meaningfully advance science and reward data sharers for meritorious contributions. The findings of this study help lay the foundation for future efforts aimed at determining the answers to these questions.

7.2 Directions for Future Research

This study represents some of the first research to undertake a comprehensive understanding of biomedical data reuse. As such, it has largely been exploratory in nature, but these findings suggest a wide range of potential avenues for future research. Some of the research directions I propose here are not currently possible, either because data sharing as described here is a new enough phenomenon that not enough historical data is yet available to conduct the analyses, or because the necessary data are simply not collected at present. I hope that the research I propose here may encourage repositories to collect the necessary data, as well as provide

direction for the development of infrastructure that will enable connections between datasets and the articles that cite them.

7.2.1 Understanding Data Requestors and Data Reuse

This study enabled a high-level understanding of who is requesting data, but it raises many additional questions about who is reusing data and patterns of reuse among requestors. For example, what accounts for the lower rate of requests among mid-career researchers, particularly associate professors, compared to early and later career researchers? Are there meaningful reasons behind the finding that associate professors are overrepresented in requests to dbGaP and underrepresented in requests to NIDDK, while the opposite is true for assistant professors?

Some of these questions could be answered by examining not only use requests, but publications arising from these requests. If systems existed to automatically connect articles to the datasets they cite, it could be possible to trace data reuse from the point of request to the point of publication, which would enable a better understanding of what various requestors are actually doing with the data. Some efforts at developing such systems are already underway. For example, the Make Data Count project aims to track data reuse by using the infrastructure that already exists to track citations to articles (Fenner et al., 2018; Make Data Count, 2019). However, tracking data reuse in this way requires not only that datasets have persistent unique identifiers that comply to a global standard, such as Digital Object Identifiers (DOIs), but also that authors know how to cite datasets and journals

correctly indicate that the citations refer to datasets. Even with the technical infrastructure in place, correct and complete tracking of dataset reuse will require significant cultural changes in science to ensure that all stakeholders in the research process document data citations in a way that enables tracking of data reuse. It is worth noting that none of the three repositories included in this study assign DOIs to their datasets, so tracking their reuse in publications is at present technically infeasible.

An even better way of finding out what requestors are doing with the data is simply asking them – since the identity of requestors is known, survey research could elicit further information about why requestors had chosen to reuse data, what they intended to do with it, what they *actually* did with it, and the impact that shared data has had on their research. This research could enable a deeper understanding of the nuances of data reuse that could inform repository plans and policies, funding decisions, and outreach to researchers.

7.2.2 Long-term Temporal Patterns

Many scientific research processes, including article citations, follow temporal patterns, and understanding these patterns can help make predictions about future performance as well as enable the development of meaningful metrics to evaluate the phenomenon in question. While this research was only able to find such patterns in requests for one of the repositories, the findings were in line with patterns observed in similar phenomena, such as article citations. With only two repositories to consider

here, it is possible that this study simply did not have enough data to draw on, so repeating this analysis with requests from other repositories could provide more meaningful results. It might also be possible to use counts of downloads or views for this analysis, in order to include repositories that do not require submission of a use request. While other parts of this study relied on use requests to understand reuse, for this analysis, that level of detail is not required, and simple annual counts of use – whether in the form of downloads, views, or requests – may be adequate.

This analysis could also yield more meaningful results if repeated again in a few years, when a longer period of request data is available. For example, the 90th percentile dbGaP datasets received more requests in the final year of available data than any previous year, so considering how the pattern of requests progresses over time could help answer some remaining questions. Will request rates continue to increase each year? It seems likely that request rates would peak and then start to decline at some point, but when will that be? Revisiting this analysis in perhaps two to five years could give a more complete picture of the temporal patterns of reuse.

The temporal analysis is also an area of research that could benefit from better connections between datasets and articles citing them. Use requests for datasets are almost certainly driven in part by the publication of articles in which researchers describe their reuse – citations to the datasets increase their visibility as well as potentially suggesting new types of reuse, when they are used in contexts that diverge from the original reason they were collected. The ability to track citations to datasets

could help explain some of the temporal patterns in requests and provide additional predictive power to models aimed at forecasting future patterns of reuse.

7.2.3 Understanding Reuse Within the Broader Research Context

Biomedical datasets are part of a complex research ecosystem that includes other research inputs and outputs, such as articles, code and software, and research funding, to name just a few. This study has provided insight into some patterns of reuse, but understanding the drivers behind those patterns likely requires looking to the broader context of how those datasets are situated within the research ecosystem.

As I have emphasized, the ability to connect datasets with the articles that cite them is crucial for understanding the context of how datasets are reused. In addition, comparing reuse of datasets by topic to the broader research funding context and the global disease burden could help provide insight into why some topics are more requested than others. These findings could identify disease areas for which datasets are under-utilized and could potentially benefit from outreach to research communities.

7.3 Conclusion

This study has provided a clearer picture of biomedical data reuse – who is reusing data, what they are doing with it, and why some datasets are more highly requested. The findings presented here demonstrate that biomedical data sharing is not a single phenomenon, but can take a range of forms that are in many cases driven by the type of data in question. Patterns of reuse differ between genomic and clinical

data, with the former being used in more meta-analyses and across a range of topics that diverges more from the original purpose for which the data were collected, while the latter tend to be reused on their own in studies that are more similar to the purpose for which they were collected. Reuse is also driven by the topic of the dataset, with more datasets covering common diseases being requested more highly than those covering rare diseases. Beyond the value of a dataset's topic in predicting the number of requests it receives, its performance early in its life is also useful in predicting how many requests it will accrue over time. Finally, data are reused by researchers from around the world and from a range of career stages, though they are in many cases most highly requested by the researchers who have the most resources with which they could collect their own data – later career researchers in the United States – as opposed to earlier career researchers and those in less-funded countries who could potentially benefit the most from having data available for reuse.

These findings are a first step in better understanding this complex phenomenon, and suggest potential avenues for future research, as well as policy and curation directions for funders and repositories. A vast amount of biomedical research data is already available, and this amount is only going to continue to grow as data sharing policies are put in place, especially when NIH eventually adopts a sharing policy that applies to all NIH funding. Understanding how those datasets are being reused is crucial to ensuring that data are shared in ways that enable meaningful reuse and that the datasets with the most value are properly curated and preserved. Many

questions still remain, but this study has taken some important first steps in better understanding data reuse.

Appendix A: Examples of Requests for Each Type of Reuse

The following table provides examples of use requests from dbGaP and NIDDK that exemplify the types of reuse described here. Request text is reproduced exactly from the original without corrections or addition of spelled-out acronyms.

Reuse Type	dbGaP Example	NIDDK Example
Original research	We propose to conduct a genome-wide scan for genetic associations with secondary phenotypes captured in the case-control sample, such as body-mass-index, lipid levels, fasting blood sugar, and serum creatinine measures using a novel secondary analysis approach. The analysis proposed represents a comprehensive and statistically rigorous genome-wide search of secondary phenotypic associations, and as such, is likely to contribute to our understanding of the underlying biologic process of peripheral arterial disease (PAD).	Cardiovascular disease is the leading cause of mortality among Hemodialysis patients. Prior research suggests that volume status and vascular stiffness are associated with cardiovascular disease. These factors are thought to be related to the rate of ultrafiltration, Hemodialysis session length, dialysate sodium concentration and phosphate intake. Though analysis of data from the HEMO Study, we seek to clarify the relationships of the relationships of these factors to one another as well as to cardiovascular outcomes among Hemodialysis patients.
Meta-analysis	The main goal of this research is to re-define the place multiple sclerosis (MS) occupies in the human disease landscape. MS is a complex autoimmune disorder of the central nervous system and is the second most common cause of neurological disability in adults after trauma. We will use de-identified genetic information from studies performed on other neurological, autoimmune, and unrelated diseases to better	Our goal of this study is to improve clinical outcomes in health. Hypertension is a topic that influences millions of lives around the world. As such, optimal targets for patients is of upmost importance. Furthermore, it is possible that optimal targets are not consistent by subpopulation groups. The NIDDK has offered access to guideline influencing studies: specifically the AASK and the MDRD trial. Our goal of

Reuse Type	dbGaP Example	NIDDK Example
	understand their similarities and differences with MS on a genome-wide scale.	our study will be to utilize advances from both these studies and pool data together to find new, meaningful clinical insights.
Comparison or control	Some children have severe seizures and other issues with their brains. Occasionally brain tissue is removed from these kids for surgical reasons. By RNA sequencing these samples we might be able to understand the cause, course and treatments for the disease. The GTex data allows us to compare these sick kids to normal individuals so that we can better understand what is going wrong in the kids.	The primary aim of this community participatory project is to conduct a translational study of the CDC Diabetes Prevention Program's successful clinic-based lifestyle intervention delivered in Community settings by community residents. Community residents at increased risk of type 2 diabetes based on BMJ, along with other risk factors, form the target population. Outcome measures include anthropometrics (e.g., BMJ, waist circumference), eating habits, and physical activity habit. The DPP data will be used to form comparison groups to examine the outcome of the community based lifestyle, intervention program.
Validation	The aim of our project is to better understand how oncogenic events cooperate during the early stages of lung cancer and during its malignant progression. To achieve this goal we are using multiple mouse models of lung cancer to study how gene gain and loss of function influences tumorigenesis. The dbGAP dataset will provide a valuable resource to help validate that recurrent genomic changes seen in our mouse models are	To date the majority of studies have focused on chronic kidney disease as a single entity with respect to outcomes. We have preliminary data to suggest that in heart failure populations this may not be correct and that the underlying pathophysiology may be highly relevant with respect to the adverse prognosis. To date we have validated these findings in 4 heart failure datasets. Interestingly, there did not appear to be any relationship between heart failure severity

Reuse Type	dbGaP Example	NIDDK Example
	relevant to the human disease. Ultimately, our goal is to identify new targets for diagnosis, prognosis and personalized treatment of patients."	and the strength of this interaction. As a result it is possible that the above described observations may not be restricted to heart failure populations and thus we are requesting the MDRD dataset to investigate these findings
Statistical methods	We are requesting the late onset Alzheimers disease data to apply the statistical methods that we develop for mapping complex genetic traits. Complex genetic traits are caused by more than one disease gene and/or non-genetic traits. Our methods take into account this fact to map the disease genes. We have developed a method that does not require disease model specification, i.e., the inheritance pattern of the disease in a family, which is unknown in real life but many methods need its specification. To study its properties, we have applied the method to simulated data. Now we need to apply it to a real data and so we are requesting this family data.	In most longitudinal medical researches, the spacing of visits is usually the same for all subjects (unbalanced design). In this study, we will evaluate how unbalanced design with increasing the frequency of visits in the high risk group will influence the precision of covariate effect estimation in interval-censored time to event data. The TN01 study used this type of unbalanced design, we will use data from this study to illustrate how this unbalanced design is beneficial in term of improving precision in risk factor estimation."
Software or tool development	The goal of this research is to create software for physician researchers which allow them to rapidly identify common genetic changes among patients suffering from the same disease. That knowledge will enable physicians to better diagnose and treat disease of all types. The real world data requested for this project will ensure that the software we develop meets	Computer simulation models would enable researchers to assess the comparative-effectiveness and cost-effectiveness of alternative strategies for the prevention and treatment of type 2 diabetes. However, due to constantly evolving treatment landscape, these models need to be repeatedly updated as new evidence becomes available to

Reuse Type	dbGaP Example	NIDDK Example
	the needs of clinical personnel.	inform their structure or input values. This project aim to update the stroke, coronary heart disease, and nephropathy sub-models in MMD by using both secondary individual-level data available through NIH repository and summary data published in the literature.
Infrastructure	The Autism Sequencing Consortium (ASC) is an organization of more than 20 research groups. The ASC seeks to collectively exploit DNA sequencing to resolve a substantial fraction of the genetic factors that contribute to Autism Spectrum Disorders (ASD). Mount Sinai School of Medicine serves as the bioinformatic Hub of the ASC. As the Hub, we store and share sequence data and call variants with ASC members, and provide them with a computing platform on which they can perform analyses. The main goal of this work is to identify rare genetic variants that associate with ASD to better understand the underlying causes of ASD.	[No requests for this use type in this repository.]
Reproducibility or reanalysis study	We wish to replicate the work of Alexandrov et al. (Nature 2013; reviewed in Martincorena Science 2015) counting the number of mutations that correspond to various mutational signatures. For this we begin with a list of mutations available through TCGA .maf files; we must then add the local genetic context for these mutations, e.g.	An analysis in the DCCT, suggested that men were at increased risk for severe hypoglycaemia. This has not been replicated in other studies. We hypothesise that gender difference is not a risk factor for severe hypoglycaemia, and that the effect found in the DCCT was a spurious result of inappropriate statistical

Reuse Type	dbGaP Example	NIDDK Example
	the preceding and following nucleotides for each single-nucleotide mutation, and this information is in the requested data from TCGA.	technique.

Appendix B: Custom Stopwords Used in LDA

This list contains stopwords that were removed from the NHLBI and NIDDK datasets descriptions for the LDA topic modeling.

background	objectives
center	outcome
conclusions	participant/participants
data	research
design	sample/samples
grant	source
individual	study/studies
measure/measures	supported

Appendix C: Topic Model Term Charts

Terms associated with topics from dbGaP LDA model.

Terms associated with topics from NHLBI LDA model.

Terms associated with topics from NIDDK LDA model.

References

Acuna, D. E., Allesina, S., & Kording, K. P. (2012). Future impact: Predicting scientific success. *Nature*, *489*(7415), 201–202. https://doi.org/10.1038/489201a

Ali-Khan, S. E., Harris, L. W., & Gold, E. R. (2017). Motivating participation in open science by examining researcher incentives. *ELife*, *6*, 1–12. https://doi.org/10.7554/eLife.29319

Ali-Khan, S. E., Jean, A., MacDonald, E., & Gold, E. R. (2018). Defining success in open science. *MNI Open Research*, *2*, 2. https://doi.org/10.12688/mniopenres.12780.1

Altman, M., & Crosas, M. (2013). The evolution of data citation: From principles to implementation. *IASSIST Quarterly*, *37*(1–4), 62–70.

Altman, M., & King, G. (2007). A proposed standard for the scholarly citation of quantitative data. *D-Lib Magazine*, *13*(3/4). Retrieved from http://www.dlib.org/dlib/march07/altman/03altman.html

Anderson, M. S., Ronning, E. A., DeVries, R., & Martinson, B. C. (2010). Extending the Mertonian norms: Scientists' subscription to norms of research. *Journal of Higher Education*, *81*(3), 612–624. https://doi.org/10.1353/jhe.0.0095

Aronson, A. R., Mork, J. G., Gay, C. W., Humphrey, S. M., & Rogers, W. J. (2004). The NLM Indexing Initiative's Medical Text Indexer. *Studies in Health Technology and Informatics*, *107*(Pt 1), 268–72. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/15360816

Arzberger, P., Schroeder, P., Beaulieu, A., Bowker, G., Casey, K., Laaksonen, L., & Moorman, D. (2004). Promoting access to public research data for scientific, economic, and social development. *Data Science Journal*, *3*(November), 135–152.

Baker, W., van den Broek, A., Camon, E., Hingamp, P., Sterk, P., Stoesser, G., & Tuli, M. A. (2000). The EMBL nucleotide sequence database. *Nucleic Acids Research*, *28*(1), 19–23. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/10592171

Begley, C. G., & Ioannidis, J. P. A. (2014). Reproducibility in science. *Circulation Research*, *116*(1), 116–126. Retrieved from http://circres.ahajournals.org/content/116/1/116.long

Belter, C. W. (2014). Measuring the value of research data: A citation analysis of oceanographic data sets. *PLoS ONE*, *9*(3). https://doi.org/10.1371/journal.pone.0092590

Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., & Wheeler, D. L. (2005). GenBank. *Nucleic Acids Research*, *33*(Database issue), D34-8. https://doi.org/10.1093/nar/gki063

Bhatt, A. (2010). Evolution of clinical research: A history before and beyond James Lind. *Perspectives in Clinical Research*, *1*(1), 6–10. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/21829774

Bierer, B. E., Crosas, M., & Pierce, H. H. (2017). Data authorship as an incentive to data sharing. *New England Journal of Medicine*, *376*(17), 1684–1687.

https://doi.org/10.1056/NEJMsb1616595

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, *3*, 993–1022. Retrieved from http://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf

Bollen, J., Van de Sompel, H., Hagberg, A., & Chute, R. (2009). A principal component analysis of 39 scientific impact measures. *PLoS ONE*, *4*(6). https://doi.org/10.1371/journal.pone.0006022

Bollen, J., Van De Sompel, H., Smith, J. A., & Luce, R. (2005). Toward alternative metrics of journal impact: A comparison of download and citation data. *Information Processing and Management*, *41*(6), 1419–1440. https://doi.org/10.1016/j.ipm.2005.03.024

Borgman, C. L. (2011). The conundrum of sharing research data. *SSRN Electronic Journal*, (1–14). https://doi.org/10.2139/ssrn.1869155

Bornmann, L., & Daniel, H.-D. (2007). What do we know about the h index? *Journal of the American Society for Information Science and Technology*, *58*(9), 1381–1385. https://doi.org/10.1002/asi

Bornmann, L., & Daniel, H.-D. (2008). What do citation counts measure ? A review of studies on citing behavior. *Journal of Documentation*, *64*(1), 45–80. https://doi.org/10.1108/00220410810844150

Bouchet-Valat, M. (2019). SnowballC: Snowball stemmers based on the C "libstemmer" UTF-8 Library. R package version 0.6.0. Retrieved from https://cran.r-project.org/package=SnowballC

Bryman, A. (2006). Integrating quantitative and qualitative research: How is it done? *Qualitative Research*, *6*(1), 97–113. https://doi.org/10.1177/1468794106058877

Burrell, Q. L. (2003). Predicting future citation behavior. *Journal of the American Society for Information Science and Technology*, *54*(5), 372–378. https://doi.org/10.1002/asi.10207

Burrell, Q. L. (2008). The publication/citation process at the micro level: A case study. *COLLNET Journal of Scientometrics and Information Management*, *3*(1), 71–77. https://doi.org/10.1080/09737766.2009.10700866

Callahan, A., Winnenburg, R., & Shah, N. H. (2018). Analysis : U-Index, a dataset and an impact metric for informatics tools and databases. *Scientific Data*, (March), 1–10.

Carpenter, C. R., Cone, D. C., & Sarli, C. C. (2014). Using publication metrics to highlight academic productivity and research impact. *Academic Emergency Medicine*, *21*(10), 1160–1172. https://doi.org/10.1111/acem.12482

Coady, S. A., Mensah, G. A., Wagner, E. L., Goldfarb, M. E., Hitchcock, D. M., & Giffen, C. A. (2017). Use of the National Heart, Lung, and Blood Institute Data Repository. *New England Journal of Medicine*, *376*(19), 1849–1858. https://doi.org/10.1056/NEJMsa1603542

CODATA-ICSTI Task Group on Data Citation Standards and Practices. (2013). Out of cite, out of mind: The current state of practice, policy, and technology for the citation of data. *Data Science Journal*, *12*(September), 1–75. https://doi.org/10.2481/dsj.OSOM13-043

Collins, F. S., Morgan, M., & Patrinos, A. (2003). The Human Genome Project: Lessons from large-scale biology. *Science*, *300*(5617), 286–290. https://doi.org/10.1126/science.1084564

Compute Canada. (2018). JupyterHub. Retrieved November 12, 2018, from https://docs.computecanada.ca/wiki/JupyterHub#cite_note-1

Consejo Superior de Investigaciones Científicas. (2019). Ranking web of universities. Retrieved March 3, 2019, from http://www.webometrics.info/en/node/54

Costello, M. J. (2009). Motivating online publication of data. *BioScience*, *59*(5), 418–427. https://doi.org/10.1525/bio.2009.59.5.9

Cozzens, S. E. (1985). Comparing the sciences: Citation context analysis of papers from neuropharmacology and the sociology of science. *Social Studies of Science*, *15*(1), 127–153. https://doi.org/10.1177/030631285015001005

Data Citation Synthesis Group. (2014). *Joint declaration of data citation principles*. (M. Martone, Ed.). FORCE11. Retrieved from https://www.force11.org/group/joint-declaration-data-citation-principles-final

de Solla Price, D. (1976). A general theory of bibliometric and other cumulative advantage processes. *Journal of the American Society for Information Science*, *27*(5–6), 292–306.

Dell EMC. (2019). Cold Data Storage. Retrieved March 17, 2019, from https://www.dellemc.com/en-us/glossary/cold-data-storage.htm

Demmer, L. A., & Waggoner, D. J. (2014). Professional medical education and genomics. *Annual Review of Genomics and Human Genetics*, *15*(1), 507–516.

https://doi.org/10.1146/annurev-genom-090413-025522

Diabetes Prevention Program Outcomes Study. (2016). Data dictionary. Retrieved from https://repository.niddk.nih.gov/media/studies/dppos/Data Dictionary.pdf

Dormann, C. F., Elith, J., Bacher, S., Buchmann, C., Carl, G., Carré, G., … Lautenbach, S. (2013). Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. *Ecography*, *36*(1), 27–46. https://doi.org/10.1111/j.1600-0587.2012.07348.x

Durieux, V., & Gevenois, P. A. (2010). Bibliometric indicators: Quality measurements of scientific publication. *Radiology*, *255*(2), 342–351. https://doi.org/10.1148/radiol.09090626

Edmunds, S. C., Pollard, T. J., Hole, B., & Basford, A. T. (2012). Adventures in data citation: Sorghum genome data exemplifies the new gold standard. *BMC Research Notes*, *5*(223). https://doi.org/10.1186/1756-0500-5-223

Edwards, M. A., & Roy, S. (2017). Academic research in the 21st century: Maintaining scientific integrity in a climate of perverse incentives and hypercompetition. *Environmental Engineering Science*, *34*(1), 51–61. https://doi.org/10.1089/ees.2016.0223

Eom, Y. H., & Fortunato, S. (2011). Characterizing and modeling citation dynamics. *PLoS ONE*, *6*(9), 1–7. https://doi.org/10.1371/journal.pone.0024926

Etikan, I., Abubakar Musa, S., & Sunusi Alkassim, R. (2016). Comparison of convenience sampling and purposive sampling. *American Journal of Theoretical and Applied Statistics*, *5*(1), 1–4. https://doi.org/10.11648/j.ajtas.20160501.11

Faniel, I. M., & Jacobsen, T. E. (2010). Reusing scientific data: How earthquake engineering researchers assess the reusability of colleagues' data. *Computer Supported Cooperative Work*, *19*(3–4), 355–375. https://doi.org/10.1007/s10606-010-9117-8

Faniel, I. M., Kriesberg, A., & Yakel, E. (2015). Social scientists' satisfaction with data reuse. *Journal of the Association for Information Science and Technology*, *67*(6), 1404–1416. https://doi.org/10.1002/asi.23480

Federer, L. (2018). Quantifying biomedical data reuse: Do citations tell the whole story? (under revision for JASIST).

Federer, L., Belter, C. W., Joubert, D. J., Livinski, A., Lu, Y.-L., Snyders, L. N., & Thompson, H. (2018). Data sharing in PLOS ONE: An analysis of Data Availability Statements. *PLoS One*, *13*(5), e0194768. https://doi.org/10.1371/journal.pone.0194768

Federer, L., Lu, Y.-L., Joubert, D. J., Welsh, J., & Brandys, B. (2015). Biomedical data sharing and reuse: Attitudes and practices of clinical and scientific research staff. *PLoS One*, *10*(6), e0129506. https://doi.org/10.1371/journal.pone.0129506

Fenner, M., Lowenberg, D., Jones, M., Needham, P., Vieglais, D., Abrams, S., … Chodacki, J. (2018). Code of Practice for Research Data Usage Metrics Release 1. *PeerJ Preprints*, 1–43. https://doi.org/10.7287/peerj.preprints.26505v1

Field, D., Amaral-Zettler, L., Cochrane, G., Cole, J. R., Dawyndt, P., Garrity, G. M., … Wooley, J. (2011). The Genomic Standards Consortium. *PLoS Biology*, *9*(6), e1001088. https://doi.org/10.1371/journal.pbio.1001088

Ford, D. Y. (2014). Segregation and the underrepresentation of Blacks and Hispanics in gifted education: Social inequality and deficit paradigms. *Roeper Review*, *36*(3), 143–154. https://doi.org/10.1080/02783193.2014.919563

Fortunato, S., Bergstrom, C. T., Börner, K., Evans, J. A., Helbing, D., Milojević, S., … Barabási, A.-L. (2018). Science of science. *Science*, *359*(6379), eaao0185. https://doi.org/10.1126/science.aao0185

Galligan, F., & Dyas-Correia, S. (2013). Altmetrics: Rethinking the way we measure. *Serials Review*, *39*, 56–61. https://doi.org/10.1016/j.serrev.2013.01.003

Gan, M., Dou, X., & Jiang, R. (2013). From ontology to semantic similarity: calculation of ontology-based semantic similarity. *The Scientific World Journal*, *2013*, 793091. https://doi.org/10.1155/2013/793091

Garfield, E. (1955). Citation indexes for science: A new dimension in documentation through association of ideas. *Science*, *122*(July), 108–11. Retrieved from http://science.sciencemag.org/content/122/3159/108

Garfield, E. (1964). Can citation indexing be automated? In *Statistical Assocation Methods for Mechanized Documentation* (Vol. 269, pp. 84–90). https://doi.org/10.1093/ije/dyl190

Garfield, E. (1982). More on the ethics of scientific publication: Abuses of authorship attribution and citation amnesia undermine the reward system of science. *Current Contents*, *30*, 5–10.

Garfield, E. (1987). Contemplating a science court: On the question of institutionalizing scientific factfinding. *The Scientist*, *1*(6), 9.

Garfield, E. (1989). Can a science court settle controversies between scientists? *Current Contents*, *28*(3–6), 189–192. Retrieved from http://garfield.library.upenn.edu/essays/v12p189y1989.pdf

Garfield, E. (1991). Bibliographic negligence: A serious transgression. *The Scientist*, *5*(23), 14. Retrieved from https://www.the-scientist.com/commentary/bibliographic-negligence-a-serious-transgression-60359

Garfield, E. (2002). Demand citation vigilance. *The Scientist*, *16*(2), 6. Retrieved from http://garfield.library.upenn.edu/papers/demandcitationvigilance012102.html

Garla, V. N., & Brandt, C. (2012). Semantic similarity in the biomedical domain: An evaluation across knowledge sources. *BMC Bioinformatics*, *13*, 261. https://doi.org/10.1186/1471-2105-13-261

Gerrish, S. M., & Blei, D. M. (2010). A language-based approach to measuring scholarly impact. In *Proceedings of the 27th International Conference on International Conference on Machine Learning* (pp. 375–382). https://doi.org/10.1002/chin.200533198

Giffen, C. A., Carroll, L. E., Adams, J. T., Brennan, S. P., Coady, S. A., & Wagner, E. L. (2015). Providing contemporary access to historical biospecimen collections: Development of the NHLBI Biologic Specimen and Data Repository Information Coordinating Center (BioLINCC). *Biopreservation and Biobanking*, *13*(4), 271–9. https://doi.org/10.1089/bio.2014.0050

Giles, J. R. A. (1995). The what, why, when, how, where and who of geological data management. *Geological Society, London, Special Publications*, *97*(1), 1–4. https://doi.org/10.1144/GSL.SP.1995.097.01.01

Ginsburg, I. (2001). The disregard syndrome, a menace to honest science. *The Scientist*, *15*(24). Retrieved from https://www.the-scientist.com/opinion-old/the-disregard-syndrome-a-menace-to-honest-science-53924

Gold, E. R., Ali-Khan, S. E., Allen, L., Ballell, L., Barral-Netto, M., Carr, D., … Thelwall, M. (2018). An open toolkit for tracking open science partnership implementation and impact. *F1000Research*, *2*. https://doi.org/10.21955/GATESOPENRES.1114891.1

Gorgolewski, K. J., Margulies, D. S., & Milham, M. P. (2013). Making data sharing count: A publication-based solution. *Frontiers in Neuroscience*, *7*, 9. https://doi.org/10.3389/fnins.2013.00009

Grüning, B., Chilton, J., Köster, J., Dale, R., Soranzo, N., van den Beek, M., … Taylor, J. (2018). Practical computational reproducibility in the life sciences. *Cell Systems*, *6*(6), 631–635. https://doi.org/10.1016/j.cels.2018.03.014

Hansson, M. G., Lochmüller, H., Riess, O., Schaefer, F., Orth, M., Rubinstein, Y., … Woods, S. (2016). The risk of re-identification versus the need to identify individuals in rare disease research. *European Journal of Human Genetics*, *24*(11), 1553–1558. https://doi.org/10.1038/ejhg.2016.52

Hazelkorn, E. (2013). How rankings are reshaping higher education. In *Rankings and the Reshaping of Higher Education: The Battle for World-Class Excellence.* (pp.

1–8). https://doi.org/10.1057/9781137446671

Henderson, T., & Kotz, D. (2015). Data citation practices in the CRAWDAD wireless network data archive. *D-Lib Magazine*, *21*(1), 1. https://doi.org/10.1045/january2015-henderson

Hey, T., Tansley, S., & Tolle, K. (2009). Jim Gray on eScience: A transformed scientific method. In *The Fourth Paradigm: Data-Intensive Scientific Discovery* (pp. xvii–xxxi). Redmond, WA: Microsoft Research. Retrieved from http://research.microsoft.com/en-us/collaboration/fourthparadigm/4th_paradigm_book_jim_gray_transcript.pdf

Hicks, D., Wouters, P., Waltman, L., de Rijcke, S., & Rafols, I. (2015). Bibliometrics: The Leiden Manifesto for research metrics. *Nature*, *520*(7548), 429–431. https://doi.org/10.1038/520429a

Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences*, *102*(46), 16569–16572. https://doi.org/10.1073/pnas.0507655102

Hirsch, J. E. (2007). Does the h index have predictive power? *Proceedings of the National Academy of Sciences*, *104*(49), 19193–19198. https://doi.org/10.1073/pnas.0707962104

Holden, G., Rosenberg, G., & Barker, K. (1994). Bibliometrics: A Potential decision making aid in hiring, reappointment, tenure and promotion decisions. *Social Work*, *39*(4), 421–431. https://doi.org/10.1300/J010v41n03

Holdren, J. P. (2013). Increasing access to the results of federally funded scientific

research. Retrieved July 19, 2017, from https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf

Hong, E. P., & Park, J. W. (2012). Sample size and statistical power calculation in genetic association studies. *Genomics & Informatics*, *10*(2), 117–22. https://doi.org/10.5808/GI.2012.10.2.117

Hopkins, P. C., Yazigi, N., & Nylund, C. M. (2017). Incidence of biliary atresia and timing of hepatoportoenterostomy in the United States. *Journal of Pediatrics*, *187*, 253–257. https://doi.org/10.1016/j.jpeds.2017.05.006

Hotho, A., Andreas, N., & Paaß, G. (2005). A brief survey of text mining. *LDV-Forum 20*, (1), 19–62. https://doi.org/10.1111/j.1365-2621.1978.tb09773.x

Human Health and Heredity in Africa. (2019). Vision. Retrieved March 21, 2019, from https://h3africa.org/index.php/about/vision/

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine*, *2*(8), e124. https://doi.org/10.1371/journal.pmed.0020124

Ioannidis, J. P. A. (2014). How to make more published research true. *PLoS Medicine*, *11*(10), e1001747. https://doi.org/10.1371/journal.pmed.1001747

Jagodnik, K. M., Koplev, S., Jenkins, S. L., Ohno-Machado, L., Paten, B., Schurer, S. C., … Ma'ayan, A. (2017). Developing a framework for digital objects in the Big Data to Knowledge (BD2K) commons: Report from the Commons Framework Pilots workshop. *Journal of Biomedical Informatics*, *71*, 49–57. https://doi.org/10.1016/J.JBI.2017.05.006

Kaiser, J. (2019). Data sharing will be a major thrust of Trump's $500 million childhood cancer plan. *Science*. https://doi.org/10.1126/science.aax1698

Kim, Y., & Yoon, A. (2017). Scientists' data reuse behaviors: A multilevel analysis. *Journal of the Association for Information Science and Technology*, *68*(12). https://doi.org/10.1002/asi.23892

Knoppers, B. M. (2014). Framework for responsible sharing of genomic and health-related data. *The HUGO Journal*, *8*(1). https://doi.org/10.1186/s11568-014-0003-1

Knoppers, B. M., Harris, J. R., Budin, I., & Edward, L. (2014). A human rights approach to an international code of conduct for genomic and clinical data sharing. *Human Genetics*, (1), 895–903. https://doi.org/10.1007/s00439-014-1432-6

Kochen, M. (1987). How well do we acknowledge intellectual debts? *Journal of Documentation*, *43*(1), 54–64. https://doi.org/10.1108/eb026801

Kotsiantis, S. B. (2007). Supervised machine learning: A review of classification techniques. *Informatica*, *31*, 249–268. Retrieved from http://www.informatica.si/index.php/informatica/article/viewFile/148/140

Laine, H. (2017). Afraid of scooping – Case study on researcher strategies against fear of scooping in the context of open science. *Data Science Journal*, *16*, 29. https://doi.org/10.5334/dsj-2017-029

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*(1), 159–74. Retrieved from

http://www.ncbi.nlm.nih.gov/pubmed/843571

Lane, J. (2010). Let's make science metrics more scientific. *Nature*, *464*(7288), 488–489. https://doi.org/10.1038/464488a

Langille, M. G. I., Ravel, J., & Fricke, W. F. (2018). "Available upon request": not good enough for microbiome data! *Microbiome*, *6*(1), 8. https://doi.org/10.1186/s40168-017-0394-z

Latour, B., & Woolgar, S. (1986). *Laboratory Life*. Princeton, NJ: Princeton University Press.

Leonelli, S. (2014). What difference does quantity make? On the epistemology of big data in biology. *Big Data & Society*, *1*(1). https://doi.org/10.1177/2053951714534395

Levin, N., & Leonelli, S. (2017). How does one "open" science? Questions of value in biological research. *Science Technology and Human Values*, *42*(2), 280–305. https://doi.org/10.1177/0162243916672071

Li, J. (2014). Citation curves of "all-elements-sleeping-beauties": "flash in the pan" first and then "delayed recognition." *Scientometrics*, *100*(2), 595–601. https://doi.org/10.1007/s11192-013-1217-z

Longo, D. L., & Drazen, J. M. (2016). Data Sharing. *New England Journal of Medicine*, *374*(3), 276–277. https://doi.org/10.1056/NEJMe1516564

Maes, M. (2015). A review on citation amnesia in depression and inflammation research. *Neuro Endocrinology Letters*, *36*(1), 1–6.

Magerman, T., van Looy, B., & Song, X. (2010). Exploring the feasibility and

accuracy of Latent Semantic Analysis based text mining techniques to detect similarity between patent documents and scientific publications. *Scientometrics*, *82*(2), 289–306. https://doi.org/10.1007/s11192-009-0046-6

Make Data Count. (2019). About. Retrieved March 23, 2019, from https://makedatacount.org/about/

Mann, G. S., Mimno, D., & McCallum, A. (2006). Bibliometric impact measures leveraging topic analysis. In *Proceedings of the 6th ACM/IEEE-CS Joint Conference on Digital Libraries - JCDL '06* (p. 65). https://doi.org/10.1145/1141753.1141765

Manolio, T. A., & Murray, M. F. (2014). The growing role of professional societies in educating clinicians in genomics. *Genetics in Medicine*, *16*(8), 571–572. https://doi.org/10.1038/gim.2014.6

Medical Library Association. (2019). Librarians without Borders. Retrieved March 21, 2019, from https://www.mlanet.org/page/librarians

Merton, R. K. (1942). The normative structure of science. In N. Storer (Ed.), *The Sociology of Science: Theoretical and Empirical Investigations* (pp. 267–278). Chicago: University of Chicago Press.

Merton, R. K. (1968). The Matthew effect in science. *Science*, *159*(3810), 56–63. Retrieved from http://www.unc.edu/~fbaum/teaching/PLSC541_Fall06/Merton_Science_1968.pdf

Merton, R. K. (1983). Foreward. In *Citation Indexing: Its Theory and Application in*

Science, Technology, and Humanities (pp. v–ix). Philadelphia: ISI Press. Retrieved from http://www.garfield.library.upenn.edu/cifwd.html

Meyer, D., Hornik, K., & Feinerer, I. (2008). Text Mining Infrastructure in R. *Journal of Statistical Software*, *25*(5), 1–54. Retrieved from http://epub.wu.ac.at/3978/%5Cnhttp://epub.wu.ac.at/%5Cnhttp://www.jstatsoft.org/

Meystre, S. M., Lovis, C., Bürkle, T., Tognola, G., Budrionis, A., & Lehmann, C. U. (2017). Clinical data reuse or secondary use: Current status and potential future progress. *Yearbook of Medical Informatics*, *26*(01), 38–52. https://doi.org/10.15265/IY-2017-007

Mitroff, I. I. (1974). Norms and counter-norms in a select group of the Apollo moon scientists: A case study of the ambivalence of scientists. *American Sociological Review*, *39*(4), 579–595. https://doi.org/10.2307/2094423

Moher, D., Naudet, F., Cristea, I. A., Miedema, F., Ioannidis, J. P. A., & Goodman, S. N. (2018). Assessing scientists for hiring, promotion, and tenure. *PLOS Biology*, *16*(3), e2004089. https://doi.org/10.1371/journal.pbio.2004089

Mole. (2004). Stealing thunder I. *Journal of Cell Science*, *117*(Pt 15), 3073–4. https://doi.org/10.1242/jcs.01281

Mooney, H., & Newton, M. (2012). The anatomy of a data citation: Discovery, reuse, and credit. *Journal of Librarianship and Scholarly Communication*, *1*(1), eP1035. https://doi.org/10.7710/2162-3309.1035

Mork, J. G., Aronson, A., & Demner-Fushman, D. (2017). 12 years on: Is the NLM

medical text indexer still useful and relevant? *Journal of Biomedical Semantics*, *8*(8). https://doi.org/10.1186/s13326-017-0113-5

Mork, J. G., Yepes, A. J. J., & Aronson, A. R. (2013). *The NLM Medical Text Indexer System for Indexing Biomedical Literature*. Retrieved from https://ii.nlm.nih.gov/Publications/Papers/MTI_System_Description_Expanded_2013_Accessible.pdf

Moura, D. C., López, M. A. G., Cunha, P., de Posada, N. G., Pollan, R. R., Ramos, I., … Fernandes, T. C. (2013). Benchmarking datasets for breast cancer Computer-Aided Diagnosis (CADx). In J. Ruiz-Shulcloper & G. Sanniti di Baja (Eds.), *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications. CIARP 2013* (pp. 326–333). Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-41822-8_41

Munafò, M. R., Nosek, B. A., Bishop, D. V. M., Button, K. S., Chambers, C. D., Percie, N., … Wagenmakers, E. (2017). A manifesto for reproducible science. *Nature Human Behavior*, *1*(January), 1–9. https://doi.org/10.1038/s41562-016-0021

Murray, M. F. (2014). Educating physicians in the era of genomic medicine. *Genome Medicine*, *6*(6), 45. https://doi.org/10.1186/gm564

National Cancer Institute. (2019). Common cancer types. Retrieved March 2, 2019, from https://www.cancer.gov/types/common-cancers

National Center for Biotechnology Information. (2018). dbGaP. Retrieved March 29, 2018, from https://www.ncbi.nlm.nih.gov/gap

National Center for Education Statistics. (2017). Digest of Education Statistics, 2017. Retrieved December 3, 2018, from https://nces.ed.gov/programs/digest/d17/tables/dt17_315.20.asp?current=yes

National Center for Health Statistics. (2017). FastStats: Leading causes of death. Retrieved March 2, 2019, from https://www.cdc.gov/nchs/fastats/leading-causes-of-death.htm

National Heart, Lung, and Blood Institute. (2008). Procedures for requesting data sets (Archived at Internet Archive). Retrieved March 1, 2019, from https://web.archive.org/web/20081120052240/http://www.nhlbi.nih.gov./resources/deca/prcdrs.htm

National Heart, Lung, and Blood Institute. (2018). BioLINCC: Biologic Specimen and Data Repository Information Coordinating Center. Retrieved March 29, 2018, from https://biolincc.nhlbi.nih.gov/home/

National Human Genome Research Institute. (2012). A Brief History of the Human Genome Project. Retrieved March 29, 2019, from https://www.genome.gov/12011239/a-brief-history-of-the-human-genome-project/

National Institute of Diabetes and Digestive and Kidney Diseases. (2018). NIDDK Central Repository. Retrieved March 29, 2018, from https://repository.niddk.nih.gov/home/

National Institutes of Health. (2016). PAR-16-256: Cancer-related Behavioral Research through Integrating Existing Data (R01). Retrieved March 12, 2019,

from https://grants.nih.gov/grants/guide/pa-files/PAR-16-256.html

National Institutes of Health. (2017a). PA-17-289: Leveraging population-based cancer registry data to study health disparities (R01). Retrieved March 12, 2019, from https://grants.nih.gov/grants/guide/pa-files/PA-17-289.html

National Institutes of Health. (2017b). Principles and guidelines for reporting preclinical research. Retrieved March 23, 2019, from https://www.nih.gov/research-training/rigor-reproducibility/principles-guidelines-reporting-preclinical-research

National Institutes of Health. (2018a). Estimates of funding for various Research, Condition, and Disease Categories (RCDC). Retrieved from https://report.nih.gov/categorical_spending.aspx

National Institutes of Health. (2018b). *NIH strategic plan for data science*. https://doi.org/10.1109/OFC.2007.4348300

National Institutes of Health. (2018c). NOT-OD-19-014: Request for Information (RFI) on proposed provisions for a draft data management and sharing policy for NIH funded or supported research. Retrieved November 11, 2018, from https://grants.nih.gov/grants/guide/notice-files/NOT-OD-19-014.html

National Institutes of Health. (2018d). PA-17-467: Secondary Analyses of Existing Alcohol Research Data (R01). Retrieved March 12, 2019, from https://grants.nih.gov/grants/guide/pa-files/PA-17-467.html

National Institutes of Health Fogarty International Center. (2019). Sub-Saharan African region information, grants and resources. Retrieved March 21, 2019,

from https://www.fic.nih.gov/WorldRegions/Pages/SubSaharanAfrica.aspx

National Institutes of Health Office of Extramural Research. (2004). Frequently asked questions (FAQs) on data sharing. Retrieved November 11, 2018, from https://grants.nih.gov/grants/policy/data_sharing/data_sharing_faqs.htm#912

National Institutes of Health Office of Extramural Research. (2016). NIH Data Sharing Policy. Retrieved July 19, 2017, from https://grants.nih.gov/grants/policy/data_sharing/

National Institutes of Health Office of Science Policy. (2017). NIH Genomic Data Sharing Policy. Retrieved July 19, 2017, from https://osp.od.nih.gov/scientific-sharing/policies/

National Institutes of Health Research Portfolio Online Reporting Tools. (2018). NIH awards by location and organization. Retrieved February 7, 2019, from https://report.nih.gov/award/index.cfm

National Library of Medicine. (2018). MeSH on Demand. Retrieved October 27, 2018, from https://meshb.nlm.nih.gov/MeSHonDemand

National Library of Medicine. (2019). Fact Sheet: MEDLINE® Journal Selection. Retrieved March 17, 2019, from https://www.nlm.nih.gov/lstrc/jsel.html

National Science Foundation. (2010). Dissemination and sharing of research results. Retrieved from http://www.nsf.gov/bfa/dias/policy/dmp.jsp

Nature Publishing Group. (2017). Availability of data & materials. Retrieved April 2, 2017, from http://www.nature.com/authors/policies/availability.html

Nonalcoholic Fatty Liver Disease (NAFLD) Adult Database. (2016). Data dictionary.

Retrieved from https://repository.niddk.nih.gov/media/studies/nafld_pediatric/Data_Dictionary.pdf

Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S. D., Breckler, S. J., … Yarkoni, T. (2015a). Promoting an open research culture. *Science*, *348*(6242), 1422 LP-1425. Retrieved from http://science.sciencemag.org/content/348/6242/1422.abstract

Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S. D., Breckler, S. J., … Yarkoni, T. (2015b). Promoting an Open research culture. *Science*, *348*(6242), 1422–1425. Retrieved from http://science.sciencemag.org/content/348/6242/1422.abstract

Nosek, B. A., & Bar-Anan, Y. (2012). Scientific Utopia: I. Opening Scientific Communication. *Psychological Inquiry*, *23*(3), 217–243. https://doi.org/10.1080/1047840X.2012.692215

Ó Conchúir, S., Barlow, K. A., Pache, R. A., Ollikainen, N., Kundert, K., O'Meara, M. J., … Kortemme, T. (2015). A web resource for standardized benchmark datasets, metrics, and Rosetta protocols for macromolecular modeling and design. *PLOS ONE*, *10*(9), e0130433. https://doi.org/10.1371/journal.pone.0130433

Olfson, M., Wall, M. M., & Blanco, C. (2017). Incentivizing data sharing and collaboration in medical research-the S-index. *JAMA Psychiatry*, *74*(1), 5–6. https://doi.org/10.1001/jamapsychiatry.2016.2610

Palevitz, B. A. (1997). The ethics of citation: A matter of science's family values. *The Scientist*. Retrieved from https://www.the-scientist.com/opinion-old/the-ethics-of-citation-a-matter-of-sciences-family-values-57456

Paltoo, D. N., Rodriguez, L. L., Feolo, M., Gillanders, E., Ramos, E. M., Rutter, J. L., … Green, E. D. (2014). Data use under the NIH GWAS Data Sharing Policy and future directions. *Nature Genetics*, *46*(9), 934–938. https://doi.org/10.1038/ng.3062

Parolo, P. D. B., Pan, R. K., Ghosh, R., Huberman, B. A., Kaski, K., & Fortunato, S. (2015). Attention decay in science. *Journal of Informetrics*, *9*(4), 734–745. https://doi.org/10.1016/J.JOI.2015.07.006

Pasquetto, I. V., Randles, B. M., & Borgman, C. L. (2017). On the reuse of scientific data. *Data Science Journal*, *16*, 1–9. https://doi.org/10.5334/dsj-2017-008

Penner, O., Pan, R. K., Petersen, A. M., Kaski, K., & Fortunato, S. (2013). On the predictability of future impact in science. *Scientific Reports*, *3*, 1–8. https://doi.org/10.1038/srep03052

Pepe, A., Goodman, A., Muench, A., Crosas, M., & Erdmann, C. (2014). How do astronomers share data? Reliability and persistence of datasets linked in AAS publications and a qualitative study of data practices among US astronomers. *PLoS ONE*, *9*(8), e104798. https://doi.org/10.1371/journal.pone.0104798

Pesquita, C., Faria, D., Falcão, A. O., Lord, P., & Couto, F. M. (2009). Semantic similarity in biomedical ontologies. *PLoS Computational Biology*, *5*(7), e1000443. https://doi.org/10.1371/journal.pcbi.1000443

Piwowar, H. A. (2010). A method to track dataset reuse in biomedicine: Filtered GEO accession numbers in PubMed Central. *Proceedings of the ASIST Annual Meeting*, *47*, 1–2. https://doi.org/10.1002/meet.14504701450

Piwowar, H. A., Becich, M. J., Bilofsky, H., Crowley, R. S., & on behalf of the caBIG Data Sharing and Intellectual Capital Workspace. (2008). Towards a data sharing culture: Recommendations for leadership from academic health centers. *PLoS Medicine*, *5*(9), e183. https://doi.org/10.1371/journal.pmed.0050183

Piwowar, H. A., Carlson, J. D., & Vision, T. J. (2011). Beginning to track 1000 datasets from public repositories into the published literature. *Proceedings of the ASIST Annual Meeting*, *48*. https://doi.org/10.1002/meet.2011.14504801337

Piwowar, H. A., Day, R. S., & Fridsma, D. B. (2007). Sharing detailed research data is associated with increased citation rate. *PLoS ONE*, *2*(3), e308. https://doi.org/10.1371/journal.pone.0000308

Popejoy, A. B., & Fullerton, S. M. (2016). Genomics is failing on diversity. *Nature*, *538*(7624), 161–164. https://doi.org/10.1038/538161a

Powledge, T. M. (2003). Revisiting Bermuda. *Genome Biology*, *4*(1). https://doi.org/10.1186/gb-spotlight-20030311-01

Priem, J. (2014). Altmetrics. In B. Cronin & C. R. Sugimoto (Eds.), *Beyond Bibliometrics: Harnessing Multidimensional Indicators of Scholarly Impact* (pp. 263–287). Cambridge, MA: MIT Press.

Pronk, T. E., Wiersma, P. H., van Weerden, A., & Schieving, F. (2015). A game theoretic analysis of research data sharing. *PeerJ*, *3*, e1242.

https://doi.org/10.7717/peerj.1242

Pryor, G. (2009). Multi-scale Data Sharing in the Life Sciences: Some Lessons for Policy Makers. *International Journal of Digital Curation*, *4*(3), 71–82. https://doi.org/10.2218/ijdc.v4i3.115

Raza, S., & Hall, A. (2017). Genomic medicine and data sharing. *British Medical Bulletin*, *123*(1), 35–45. https://doi.org/10.1093/bmb/ldx024

Reality check on reproducibility. (2016). *Nature*, *533*, 437.

Richesson, R. L., & Nadkarni, P. (2011). Data standards for clinical research data collection forms: Current status and challenges. *Journal of the American Medical Informatics Association*, *18*(3), 341–346. https://doi.org/10.1136/amiajnl-2011-000107

Rinker, T. W. (2018). {textstem}: Tools for stemming and lemmatizing text. R package version 0.1.4. Retrieved from http://github.com/trinker/textstem

Robinson-García, N., Jiménez-Contreras, E., & Torres-Salinas, D. (2015). Analyzing data citation practices using the Data Citation Index. *Journal of the American Society for Information Science and Technology*, *18071*, 12. https://doi.org/10.1002/asi.23529

Rolland, B., & Lee, C. P. (2013). Beyond trust and reliability: reusing data in collaborative cancer epidemiology research. In *Proceedings of the ACM Conference on Computer-Supported Cooperative Work and Social Computing (CSCW'13)* (pp. 435–444). https://doi.org/10.1145/2441776.2441826

Savage, C. J., Vickers, A. J., Kats, J., & Molenaar, D. (2009). Empirical study of data

sharing by authors publishing in PLoS journals. *PLoS ONE*, *4*(9), e7078. https://doi.org/10.1371/journal.pone.0007078

Schlögl, C., Gorraiz, J., Gumpenberger, C., Jack, K., & Kraker, P. (2014). Comparison of downloads, citations and readership data for two information systems journals. *Scientometrics*, *101*(2), 1113–1128. https://doi.org/10.1007/s11192-014-1365-9

Serwadda, D., Ndebele, P., Grabowski, M. K., Bajunirwe, F., & Wanyenze, R. K. (2018). Open data sharing and the Global South: Who benefits? *Science*, *359*(6376), 642–643. https://doi.org/10.1126/science.aap8395

Sheehan, J., Hirschfeld, S., Foster, E., Ghitza, U., Goetz, K., Karpinski, J., … Huerta, M. (2016). Improving the value of clinical research through the use of Common Data Elements. *Clinical Trials (London, England)*, *13*(6), 671–676. https://doi.org/10.1177/1740774516653238

Silge, J., & Robinson, D. (2018). Topic modeling. In *Text Mining with R: A Tidy Approach*. Boston: O'Reilly Media.

Silva, L. (2014). PLOS' new data policy: Public access to data. Retrieved April 2, 2017, from http://blogs.plos.org/everyone/2014/02/24/plos-new-data-policy-public-access-data-2/

Silvello, G. (2017). Theory and practice of data citation. *Journal of the Association for Information Science and Technology*, *69*(1), 6–20. https://doi.org/10.1088/0305-4624/9/3/409

Stodden, V., Bailey, D., Borwein, J., LeVeque, R. J., Rider, W., & Stein, W. (2012).

Setting the default to reproducible: Reproducibility in computational and experimental mathematics. In *Reproducibility in Computational and Experimental Mathematics* (pp. 1–19). Retrieved from http://icerm.brown.edu/video_archive,

Stodden, V., Seiler, J., & Ma, Z. (2018). An empirical analysis of journal policy effectiveness for computational reproducibility. *Proceedings of the National Academy of Sciences*, *115*(11), 2584–2589. https://doi.org/10.1073/pnas.1708290115

Taichman, D. B., Sahni, P., Pinborg, A., Peiperl, L., Laine, C., James, A., … Backus, J. (2017). Data sharing statements for clinical trials: A requirement of the International Committee of Medical Journal Editors. *New England Journal of Medicine*, *376*(23), 2277–2279. https://doi.org/10.1056/NEJMe1705439

Tausczik, Y. R. (2016). Citation and attribution in open science: A case study. In *Proceedings of the Conference on Computer Supported Cooperative Work and Social Computing (CSCW)* (pp. 1524–1534). https://doi.org/10.1145/2818048.2820070

Tenopir, C., Allard, S., Douglass, K., Aydinoglu, A. U., Wu, L., Read, E., … Frame, M. (2011). Data sharing by scientists: practices and perceptions. *PLOS ONE*, *6*(6), e21101. https://doi.org/10.1371/journal.pone.0021101

Tenopir, C., Dalton, E. D., Allard, S., Frame, M., Pjesivac, I., Birch, B., … Dorsett, K. (2015). Changes in data sharing and data reuse practices and perceptions among scientists worldwide. *PLOS ONE*, *10*(8), e0134826.

https://doi.org/10.1371/journal.pone.0134826

The International Consortium of Investigators for Fairness in Trial Data Sharing. (2016). Toward Fairness in Data Sharing. *New England Journal of Medicine*, *375*(5), 405–407. https://doi.org/10.1056/NEJMp1605654

Thelwall, M., Haustein, S., Larivière, V., & Sugimoto, C. R. (2013). Do altmetrics work? Twitter and ten other social web services. *PLoS ONE*, *8*(5), 1–7. https://doi.org/10.1371/journal.pone.0064841

Thygesen, L. C., & Ersbøll, A. K. (2014). When the entire population is the sample: Strengths and limitations in register-based epidemiology. *European Journal of Epidemiology*, *29*(8), 551–558. https://doi.org/10.1007/s10654-013-9873-0

van Raan, A. F. J. (2004). Sleeping Beauties in science. *Budapest Scientometrics*, *59*(3), 467–472. Retrieved from https://link.springer.com/content/pdf/10.1023/B:SCIE.0000018543.82441.f1.pdf

van Raan, A. F. J. (2005). Measurement of central aspects of scientific research: Performance, interdisciplinarity, structure. *Measurement*, *3*(1), 1–19. https://doi.org/10.1207/s15366359mea0301_1

Wallis, J. C., Rolando, E., & Borgman, C. L. (2013). If we share data, will anyone use them? Data sharing and reuse in the long tail of science and technology. *PLoS ONE*, *8*(7), e67332. https://doi.org/10.1371/journal.pone.0067332

Wan, Z., Vorobeychik, Y., Xia, W., Clayton, E. W., Kantarcioglu, M., & Malin, B. (2017). Expanding access to large-scale genomic data while promoting privacy: A game theoretic approach. *American Journal of Human Genetics*, *100*(2), 316–

322. https://doi.org/10.1016/j.ajhg.2016.12.002

Wang, J. (2013). Citation time window choice for research impact evaluation. *Scientometrics*, *94*(3), 851–872. https://doi.org/10.1007/s11192-012-0775-9

Werner, R. (2015). The focus on bibliometrics makes papers less useful. *Nature*, *517*(7534), 245. https://doi.org/10.1038/517245a

Weymann, D., Laskin, J., Roscoe, R., Schrader, K. A., Chia, S., Yip, S., … Regier, D. A. (2017). The cost and cost trajectory of whole-genome analysis guiding treatment of patients with advanced cancers. *Molecular Genetics and Genomic Medicine*, *5*(3), 251–260. https://doi.org/10.1002/mgg3.281

Wickham, H. (2016). rvest: Easily harvest (scrape) web pages. R package version 0.3.2. Retrieved from https://github.com/hadley/rvest

Wickham, H. (2017a). httr: Tools for working with URLs and HTTP. R package version 1.3.1. Retrieved from https://github.com/r-lib/httr

Wickham, H. (2017b). tidyverse: Easily Install and Load "Tidyverse" Packages. R package version 1.1.1. Retrieved from https://cran.r-project.org/package=tidyverse

Wikipedia. (2018). List of academic ranks. Retrieved October 28, 2018, from https://en.wikipedia.org/wiki/List_of_academic_ranks

World Health Organization. (2019). About Hinari. Retrieved March 21, 2019, from https://www.who.int/hinari/about/en/

Yakel, E., Faniel, I. M., Kriesberg, A., & Yoon, A. (2013). Trust in digital repositories. *International Journal of Digital Curation*, *8*(1), 143–156.

https://doi.org/10.2218/ijdc.v8i1.251

Yoon, A. (2014). End users' trust in data repositories: Definition and influences on trust development. *Archival Science*, *14*(1), 17–34. https://doi.org/10.1007/s10502-013-9207-8

Yoon, A. (2017). Role of communication in data reuse. In *Proceedings of the Association for Information Science and Technology* (Vol. 54, pp. 463–471). Crystal City, VA. https://doi.org/10.1002/pra2.2017.14505401050

Zhang, Q., Cheng, Q., Huang, Y., & Lu, W. (2016). A bootstrapping-based method to automatically identify data-usage statements in publications. *Journal of Data and Information Science*, *1*(1), 1–17. https://doi.org/10.20309/jdis.201606

Zhang, Y., Jin, R., & Zhou, Z. H. (2010). Understanding bag-of-words model: A statistical framework. *International Journal of Machine Learning and Cybernetics*, *1*(1–4), 43–52. https://doi.org/10.1007/s13042-010-0001-0

Zhao, M., Yan, E., & Li, K. (2017). Data set mentions and citations: A content analysis of full-text publications. *Journal of the Association for Information Science and Technology*, *69*(1), 32–46. https://doi.org/10.1002/asi.23919

Zhou, J., & Shui, Y. (2015). MeSHSim: MeSH(Medical Subject Headings) semantic similarity measures. R package version 1.2.0. Retrieved from https://github.com/JingZhou2015/MeSHSim

Zhou, J., Shui, Y., Peng, S., Li, X., Mamitsuka, H., & Zhu, S. (2015). MeSHSim: An R/Bioconductor package for measuring semantic similarity over MeSH headings and MEDLINE documents. *Journal of Bioinformatics and Computational*

Biology, *13*(06), 1542002. https://doi.org/10.1142/S0219720015420020

Zimmerman, A. (2007). Not by metadata alone: The use of diverse forms of knowledge to locate data for reuse. *International Journal on Digital Libraries*, *7*(1–2), 5–16. https://doi.org/10.1007/s00799-007-0015-8