Does Management Matter in Scientific Laboratories? Evidence from Harvard Medical School

by

Florian Hillen
B.Sc., Technische Universität München (2017)

Submitted to the Institute for Data, Systems, and Society
and
Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degrees of
Master of Science in Technology and Policy
and
Master of Science in Electrical Engineering and Computer Science
at the
MASSACHUSETTS INSTITUTE OF TECHNOLOGY
June 2018

Signature redacted

Author ....
Institute for Data, Systems, and Society
Department of Electrical Engineering and Computer Science
May 23, 2018

Signature redacted

Certified by ....
Eric von Hippel
Professor of Management of Innovation and Engineering Systems,
MIT Sloan School of Management
Thesis Supervisor

Signature redacted

Certified by ....
Karim R. Lakhani
Professor of Business Administration, Harvard Business School
Thesis Supervisor

Signature redacted

Certified by ....
Caroline Uhler
Assistant Professor, Department of Electrical Engineering and Computer Science
Assistant Professor, Institute for Data, Systems, and Society
Thesis Reader

Accepted by ....Signature redacted....................
Munther A. Dahleh
Professor, Department of Electrical Engineering and Computer Science
Director, Institute for Data, Systems, and Society

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

Does Management Matter in Scientific Laboratories? Evidence from Harvard Medical School

by

Florian Hillen

Submitted to the Institute for Data, Systems, and Society
and the Department of Electrical Engineering and Computer Science
on May 23, 2018, in partial fulfillment of the
requirements for the degrees of
Master of Science in Technology and Policy

and

Master of Science in Electrical Engineering and Computer Science

Abstract

The high quality of modern medical care is built upon the creation of scientific knowledge generated from medical research. While the role of management practices has been rigorously explored across various industries, little is known about management in medical research. I collected data surveying principal investigators of medical research laboratories at the Harvard Medical School to examine the relationship of management practices and research outputs. I find that principal investigators with more effective management practices are associated with higher-impact research (measured by citations). This effect is stronger and more significant in younger compared to older laboratories and remains robust after using different controls. This study helps to increase the understanding of management in a scientific setting and should start a new discussion about the relevance of management in medical research.

Thesis Supervisor: Eric von Hippel
Title: Professor of Management of Innovation and Engineering Systems,
MIT Sloan School of Management

Thesis Supervisor: Karim R. Lakhani
Title: Professor of Business Administration, Harvard Business School

Thesis Reader: Caroline Uhler
Title: Assistant Professor, Department of Electrical Engineering and Computer Science
Assistant Professor, Institute for Data, Systems, and Society

Acknowledgments

This thesis embodies the completions of two of the most instructive, intense and inspiring years of my life. It could not have been possible without the support of many amazing individuals.

First of all, I want to thank Professor Karim R. Lakhani who supported me since we first met in Fall of 2015. He continuously inspires me with his morals, sense of fairness, creativity and commitment to rigorous, novel and impactful research. I am fortunate to have a mentor who invested so much time, resources and goodwill into my academic and personal growth. He is my mentor, colleague and friend.

My sincere gratitude also goes to Professor Raffaella Sadun who is a great mentor and friend. She always had an open door and thought me the essence of great research. Furthermore, I want to thank Professor Eric von Hippel, Professor Caroline Uhler and Professor Eva Guinan for being my mentors throughout my academic research at MIT. Special gratitude also goes to Aravind Subramanian and the CMap Team at the Broad Institute for their ongoing support throughout my studies.

Many new and old friends supported, challenged, taught and inspired me throughout my time at MIT and the creation of this thesis. Especially, I want to thank my roommates Andres, Alejandro and Bjarke for all the late night discussions, the mutual support, the travels and fun we have shared. Furthermore, I want to thank TPP, a program truly caring for its students, and my friends in it such as Christoph, Othmane, and Erik, for making my time so enjoyable. Also a special thanks goes to my colleagues Michael and Jin at LISH who always wanted the best for me as well as Maxi whose feedback I always highly value and appreciate. A thank you also goes to all my fellow MIT students and faculties who have worked with me throughout my two years at this fantastic university. I had the privilege to work with some of the brightest, most kind and helpful people from everywhere in the world - a truly humbling experience.

Finally, none of this would have been possible without the unconditional support of my mother, father and brother. Their advice, open ears, and support mean the world to me.
Thank you.

THIS PAGE INTENTIONALLY LEFT BLANK

Contents

THIS PAGE INTENTIONALLY LEFT BLANK

List of Figures

THIS PAGE INTENTIONALLY LEFT BLANK

List of Tables

THIS PAGE INTENTIONALLY LEFT BLANK

Chapter 1

Introduction

Medical research is the foundation of societal health care, shaping the fundamental understanding of our nature and developing future treatments for previously incurable diseases (Gostin et al., 2009). Beyond the obvious advancements in health, academic research is a large industry by itself. Every year, the federal government invests close to $40 billion into medical research at academic institutions, most of which comes from the National Institute of Health (NIH) and the National Scientific Foundation (America, 2017). It is estimated that every dollar invested by the NIH yields a return of $2.21 in economic output within one year (Macilwain, 2010). However, the advances in medical research seem to be declining in pace as the costs per approved drug increased significantly over the past years (DiMasi et al., 2016). One reason for this development is that the more researchers advance in science, the more complex future research questions become. Modern medical research becomes increasingly challenging, demanding a variety of skill sets, large resource investments and overall, more interdisciplinary and inter-institutional collaboration. But are medical research laboratories well prepared for this change?

Perhaps not. Research in general and medical research in particular is in a crisis with an increasing number PhD students reporting burn-out, an increasingly competitive job market and a replication crisis. PhD students are the life-blood and the main work force in medical research laboratories (labs), yet they feel an increasing anxiety about their work and future. A study from Nature surveying over 5,700 PhD students found out that over 25% of respondents have mental health issues and 45% of them seek mental health against depression or anxiety (Woolston, 2017). Almost one fifth of PhD students did not feel supported at their home institutions when it comes to career planning and around 30% of respondents disagreed or strongly disagreed that they receive helpful advice from their academic supervisor. Mentoring their PhD students for their future career is one of the central tasks of principle investigators (PIs) in academic labs. Yet, it seems that that they, at least to some part, fail at it. Another crisis medical research is currently facing is the replication of results (or the lack thereof). Reproducibility of experiments and results plays an imperative role in academic research. Yet, the past has shown that many scientific studies being published in well-known journals are hard, if not impossible, to replicate. It is estimated that around 50% of pre-clinical research studies are not reproducible. According to the NIH, part of the reason is a lack of training and rigor of young scientists - both responsibilities of the PI and the PhD institution. To counteract this, the NIH has initialized new educational efforts to train scientists proper conduct of research (Collins and Tabak, 2014). Additional reasons for the replication crisis are the unprecedented rate of new data which can be used for biomedical research as well as the significant pressure in academia to 'publish or perish' (Begley and Ioannidis, 2015). The pressure to publish or perish in academia and medical research in particular might also lead to other negative consequences. Many research grants, scholarships and awards are given on basis of the applicant's research productivity. Thus, researcher try to scramble together whatever they have for a publication to meet an externally set deadline instead of taking time to conduct rigorous, high-impact and novel research. In the current system and in research labs without a principal investigator managing to teach their PhD students how to best deal with this high-pressure academic world, many potentially high-impact and novel ideas can be lost - pushed aside for the supposedly clearer and incremental research projects. Thus, there are manifold challenges medical research is facing and to overcome them, there is not only a need for systematic changes in the way the grant and publication system is designed, but also in the management of medical research labs. The success of research and to overcome the above mentioned problems rely to a significant extent on the capabilities of the PI and the structure of every individual research lab.

In existing structures of medical research, the laboratories depend to a large extend on the principal investigator, who provides guidance to his fellow researchers, manages projects and resources of the lab as well as sets the general research strategy. It is fair to say that a substantial part of their job is management. When being promoted to the position of a PI, their role changes from being a specialized researcher to a manager of a complex organization. However, in general, the PI does not obtain any kind of managerial training. This lack of management education in medical research might be rooted in different reasons. First, it is plausible that PIs are unaware of the utility of management in general. Most PIs have had little contact with any managerial education during their career. Furthermore, "management" seems to have a negative connotation in the scientific community, being more associated with the opposite of having a "pure scientific motivation and rigor". Second, until now, very little is known about how PIs can manage their labs most effectively. While ineffective leadership and processes might not negate good science (Sapienza, 2004), the question is: *Would better management practices produce better science?* While the impact of good management practices has been rigorously analyzed across industry sectors (Bloom, 2007), little is known about the role of management in medical research. Existing research is very limited towards using high level data to make rather uninformative assumptions about concepts related to management. For instance, using citation count and number of authors on a paper to better understand the importance of collaborations (Wuchty et al., 2007). The current research lacks a tools to measure detailed data within the research lab about the relationship between actual management practices and scientific success. The dilemma in research in general is that management on the one hand ensures coordination and effective allocation of resources, but, on the other hand, it might limit the freedom for creativity and preferences of autonomy of the researcher. There seems to be a general uncertainty of how to deal and use management practices within medical research. As Professor Sargent joked in an interview with Nature "You might think that after 20 years I have this [how to run his lab] completely figured out, but it's still an evolving process."(Woolston, 2017). The common ground across all interviews for this study was that there is no formal training on how to run a research lab. Most PIs simply copy the practices thought by their previous supervisor without a clear feedback loop to measure if they are effective or not. Many of the interviewees in this study, considered the top tier academics in their respective fields, report a strong uncertainty what the best practices might be.

In fact, there are no known best practices regarding how to manage medical research labs. Yet, the current challenges medical research faces might be, to some extend, grounded in this lack of know-how. In order to overcome and meet future demands and to continue to make important discoveries positively impacting society's health care, it is essential to better understand how medical research labs can and should be managed. This thesis is a first important step towards creating a better understanding and shape future policy decisions.

This study uses an adapted version of the World Management Survey (e.g., Bloom and Van Reenen (2007); Bloom et al. (2012c, 2010)) specified to measure management practices and organizational structures in medical research laboratories. In total, I have collected data from 133 interviews with principal investigators of research labs at Harvard Medical School. By linking this management data to scientific output information from publications, citations and funding, I address three questions: First, does management matter in medical research? Second, what drives management quality in research labs? Third, do collaborative lab structures matter for medical research outcomes? The results of this study provide suggestive evidence that management does matter in medical research. Better management practices are associated with more citations per publication, a widely-recognized proxy for research quality. Furthermore, it seems that neither lab age, size or gender of the PI drive differences in management practices. Lastly, no conclusion can be derived from the comparison of collaborative structures in research labs. The data, however, suggests that collaborative labs seem to have more publication, citation and a slightly higher management score.

This thesis will first offer a coherent overview of the systematic structures of medical research and on the current literature on management practices in medical research. I will further explain the study data and methods used for this study. Specifically, this section will explain the employed survey tool, the interview technique and the externally gathered objective outcome variables for each research lab. In the next section, clustering, I lay out an in-depth cluster analysis in order to group the sampled research labs together on basis of their actual research and use the resulting clusters as control variables in the consecutive analyses. In the chapter, results, this thesis will lay out the statistical analyses conducted to further explore my sampled research labs and finding answers to the main questions of this thesis. After the discussion of the presented results, this thesis will summarize a conclusion of the current work and implications for future efforts.

THIS PAGE INTENTIONALLY LEFT BLANK

Chapter 2

Review of Literature

2.1 Medical research - a systematic overview

2.1.1 The structure of biomedical research

A medical research lab is generally structured in a hierarchical system. It is led by an independent researcher, the principal investigator (PI), who is usually at the same time an appointed faculty member at a university (i.e. Assistant, Associate or Full Professor). The PI is responsible for setting the general research agenda of the lab, in charge of its funding and can decide all personal related questions, e.g. whom to hire and fire. Each lab has a different composition of scientists but usually a lab consists of PhD students, who are in the process of learning how to conduct high-quality scientific research on their own, and/or postdocs, who are in the process of going on the job market and become the PI of their own research lab. Additionally, most research labs have research technicians who know how to operate and conduct the scientific equipment and help with the execution of research projects. Depending on the size, funding and the PI's preferences some research labs furthermore have research scientists (full time researcher not on a tenure track career path), administrative assistants (exemplary responsibilities include scheduling meetings, rooms and appointments) and lab managers. A lab manager's responsibilities are to organize and maintain the lab's resources, such as ordering supplies and taking care of potential animal stock etc. Overall, most individuals in the research lab, especially the PhD students and postdocs, have a limited educational time within the research lab before moving on to another or their own lab. Unlike in industry, there is usually no clear career path within the organization/lab they are in. Instead, PhDs and postdocs need to leave the research lab to make the next career step. Most of the information in this paragraph comes from the book "Managing Scientists' (Sapienza, 2004).

2.1.2 The success metric of biomedical research

The success metric of biomedical research are most of high-impact publications. An academic career depends significantly on the number of publications, the journals published to and other factors, such as securing significant funding or having some sort of prominent exposure (Baruch and Hall, 2004). In general, academic productivity of a researcher is often measured by the total number of papers a researcher has published (Barnett et al., 1998). Thus, taking the number of academic publications from the researchers in a lab seems to be a legitimate measurement of the lab's productivity. But there are many factors playing into this simple definition of lab's productivity. A study from 1992 early on examined the research productivity with regards to department size (Golden and Carstensen, 1992). The findings suggest that department size is significantly correlated with number of publications. However, controlling for research support and the faculty rating of the department, this relationship becomes insignificant. This suggests that the faculty, usually also the principal investigator, plays a crucial role in the publication count. Perhaps PIs with higher rankings are simply more productive. Additionally, unobserved factors could play in, such as the prestige and exposure the PI adds to a publication when appearing as a co-author.

In order to evaluate the quality and/or impact of a publication, often used proxies include the citation count of the publication, the impact factor of the publishing journal or, to measure the quality of the author, the Hirsch-index. The most straight forward and often-used proxy for quality is the citation count of the publication. In 1990 already, Egghe and Rousseau formulated four assumptions on how citation counts are associated with the contribution of a publication (Egghe and Rousseau, 1990). First, the citation implies that the original publication is actually used by another researcher, second the citation reflects the merit of the publication, third, the reference will be made to the best possible work in the related area and fourth, the content of the cited and citing publications are related. Following this logic, the citation count can be regarded as a measurement for the impact and contribution a publication has on the respective field. Nevertheless, the more citations have been used as a proxy for research quality, the more it has been questioned in the past as well (Nieminen et al., 2006). For instance, a comparative study of 448 research paper by Nieminen et al. did not find a significant association of the quality of reporting and the statistical analysis with the number of citations in a publication. A further review of studies by Bornmann and Daniel in 2008 (Bornmann and Daniel, 2008), analyzed about 40 publications trying to understand the relationship of citations with the quality and impact of the research. They find that citing behavior is not restricted to "acknowledge intellectual and cognitive influence of colleagues scientists" but also other non-scientific factors play a role in citing a paper. Nevertheless, they conclude that citation count is still a reliable method to mesure impact of a publication. Overall, the citation count has been cautiously considered, given its problems: self-citations, the general increase of citations numbers and the correlation between the number of authors of an article and the number of citations it receives. Thus, many current research efforts are aiming to find better fitting quality metrics such as the Becker Model (Sarli et al., 2010). However, none of them having a clear answer to the downsides of citations counts. As of now, citation counts seem still to be the best proxy to measure the impact of a publication within a respective research field.

Another measurement used to determine the scientific quality is the impact factor of a journal a manuscript was published in. According to Eugene Garfield, the creator of the "impact factor", it is the measurement of how often an 'average article' in a journal has been cited in a particular year (Garfield et al., 1994). However, since its creation, it has often been misconceived as some perfect measure to determine the "true impact" of research (Hecht et al., 1998; Dong et al., 2005) instead of being simply a time-specific index for the citation rate of a journal. To some sense, it represents a citation count on the journal instead of the single publication level. A similar analogy can be drawn for the Hirsch-index as it is in some way a citation count on the individual resaearcher lebel. J.E. Hirsch introduced the Hirsch index in 2005 (Hirsch, 2005). It is defined as the number of papers, h, with citations equal or great than h. That is, if an individual researcher has an H-index of e.g. 25, she has published 25 manuscripts with 25 or more citations.

Every citation-based measure (i.e. citation count, impact factor and h-index), has to be seen in the context of many different factors playing into the citation frequency in the first place. For instance, a cohort study in 2007 comparing citation patterns in three highly considered medical journals (i.e. JAMA, Lancet and the New England Journal of Medicine) found that publications with group authorship, industry funding, and industry-favouring results are cited more often (Kulkarni et al., 2007). Furthermore, this study found that fields such as oncology and cardiology were also associated with a higher citation frequency. This is no exception, especially when comparing citation patterns across fields within medical research, the size of addressed audience and citation patterns have large variations. Comparing three large medical research fields of Cardiac & cardiovascular systems, Clinical neurology, and Surgery has shown significant difference in citation counts (Radicchi et al., 2008) across these fields. Thus, many more sophisticated bibliometric indicators try to control for the field differences but fail to take within-field heterogeneity into account. As a consequence, a study found that the citation counts of clinical intervention research is significantly underestimating its impact compared to citation counts in basic research within the same field (Van Eck et al., 2013). Overall, for each citation-based measurement there seems to be a large variance of citation counts across different research fields as well as different kind of research, i.e. clinical versus basic science.

Another important aspect of research is funding. In 2009, U.S. universities spent almost $55 billion on research coming from various sources: The federal government (59.3 %), the universities themselves (20.4%), state and local governments (6.6%), industry (5.8%) and other sources (7.9%), such as private foundations e.g. the Gates Foundation (Sapienza, 2004). The largest donors for academic research are the federal government, leading with the two agencies, the National Institute of Health (NIH) and the National Science Foundation (NSF). In 2017, the NIH spend nearly $37.3 billion in medical research (NIHReporter, 2017) and the NSF$7.5 billion (NSF, 2017). The majority of this funding, e.g. 80% of the NIH grants, is awarded through a competitive process of over 300,000 researcher applying for 50,000 grants (NIHReporter, 2017). This process should ensure the quality and novelty of funded research projects. However, it has also been strongly criticized over the past years as being imperfect and favouring incremental over novel research (Brainard, 2007). Furthermore, a study from 2011 found that the competitive process of NIH, is subjective to biases such as ethnicity or race with black applicants having a 10% lower chance of received a grant controlling for comparable credentials (Ginther et al., 2011). Funding in general is a chicken-or-egg question: Does better research lead to more funding or does more funding lead to better research? It is likely to be a combination of both. A study from 2013 analyzing the relationship between NIH funding and the H-index of the individual research found a significant correlation between them both (Svider et al., 2013). The paper concludes that the H-index could be a predictor for the chances of funding success. As one of the criteria when assessing funding applications is about the applicants and their track record, it is obvious that the applicant's work of research and its recognition, described to some part in the H-index, influences the chances of success.

While there is a mountain of literature on the advantages and disadvantages of the competitive funding process, it is imperative for the applicant, usually the principle investigator of the lab, to know how to navigate it. Funding is essential on conducting medical research as there is a clear link between the level of funding in a given area and its productivity in terms of field advancements and innovations (Moses et al., 2005). This importance of funding is more likely to increase in importance and competitiveness given that modern biomedical research becomes more complex and resource intensive (Stephan, 2012).

This section looks into the various measurements of scientific output. Given the status quo of literature, this study can draw the following conclusion. First, in order to measure the productivity of a research lab, a reasonable measurement is the count of publications, in which the last author is the principle investigator. The last authorship is usually given to the principle investigator in whose research lab the research has been conducted. Thus, it enables this study to count all publications generated by one research lab. The share of publications in which the last author is not the PI of the associated research lab or that has several last authors (as the work was a collaboration between research labs) is estimated to be negligibly small. Second, in order to approximate for the quality of lab's research output, this study concludes to use a simple citation count per publication coming out of the research lab (i.e. having the PI as a last author). Screening the relevant literature, it still seems to be the best possible estimate to measure the impact a research lab has on their respective fields. The impact factor and the H-index are both measurements better suitable to measure either the impact of a potential journal or of an individual researcher. Furthermore, the literature review reveals that in this consecutive analysis, it is important to control somehow for the research field as also the simple publication and citation count varies significantly across sub-fields. Third, this study concludes to have as a third outcome variable of the labs success the amount of funding it has. Funding is an essential part of the success of a research lab. This works two ways. Having enough funding enables the lab to conduct high-profile research without being constrained by financial resources and secondly, through the competitive NIH process, it represents additional productivity and quality metrics screened by the grant reviewers. Overall, this study uses as the three main outcome variables of research labs, number of publications, number of citations per publication and funding per year to approximate the productivity, quality and opportunities a medical research laboratory has.

2.2 Status quo of management research

The relevance of management has been analyzed rigorously across a variety of sectors ranging from corporate industry (Bloom et al., 2007; Bloom and Van Reenen, 2007) over education (Bloom et al., 2015) to health care and hospital management (West, 2001). The overall consensus is that management does matter in these industries and either improves productivity or industry-specific quality criteria. For instance, a field experiment on Indian textile firms with the treatment of free management practice consulting showed that adopting management practices raised the productivity by 17% (Bloom et al., 2013). This effect was mainly achieved by an increase of efficiency and reduced inventory. Thus, in general research there seems to be a mountain of evidence suggesting the effectiveness of better management practices.

In contrast, medical research has fundamentally different characteristics than corporate industry. Incentives, funding, objectives and bureaucracy are very different in academic research. For example, management in industry is usually aligned with well defined objectives, such as increasing the return on investment, launch a new product or increase the market share. Furthermore, employees of corporations usually have well defined career ladders and the management is incentivized to train and retain their top talent. On the contrary, PhD and postdocs in medical research labs are only in the lab for a maximum of six years after which it is considered to be a success if they become the principle investigator of their own lab. Thus, the definition of successful human resource and good management in general seems to be very different.

A better analog of medical research lab management seems to be RD organization management. Previous research on RD units has shown that management practices can increase the quality of the final product, reduce development costs and the strengthen the corporation's competitive advantage (Nobelius, 2004). Still, the characteristics of for-profit, applied research units is still significantly different than academic medical research. Thus, this study will continue to analyze the limited body of literature specifically about the topic of management and related areas in medical research.

A study looking at the incentives of academic research to patent showed that four different factors play a significant role in increasing the technology transfer: great rewards for the faculty to transfer technology, university location, university mission in support of technology transfer and the experience of the university in doing so (Friedman and Silberman, 2003). Another aspect analyzed in previous literature is the relationship between organizational structures and research productivity (Carayol and Matt, 2004). Carayol and Matt analyzed research characteristics and their relationship with two major outcome variables, patents and publications. They conclude that the organizational mix of a lab, such as having a combination of full-time researchers and university professors, is associated with productive labs. Productive labs (in terms of number of publications) also tended to patent more.

The size of the lab, individual promotions and the role of non-permanent researchers additionally play a big role in having a productive research labs. A third aspect which has been investigated extensively in the past is the role of collaboration on scientific output (Diamond, 1985; Wuchty et al., 2007; Chinchilla-Rodríguez et al., 2012), which has the general consensus that greater collaboration (e.g. measured by number of co-authors) is associated with a great number of citations or higher impact factor of publishing journal. Part of this might be explained by a larger exposure of the research paper. The co-authors of these papers are more likely to have more collaborative projects over the course of their research career and re-cite their fellow co-authors more often (Gazni and Thelwall, 2014). Another reason comes from the field of psychology. Studies examining the dynamics of groups showed that groups with high diversity and not conformity are more likely to produce more novel and higher quality research (De Dreu and West, 2001). A case study on academic research laboratories in Thailand concluded that collaboration provides greater access to research knowledge. Furthermore, trust of collaborating parties seems to be detrimental to gain mutual benefits and Information and Communication Technologies (ICTs) are essential for successful collaborative projects (Numprasertchai and Igel, 2005). An extensive study of over 19 million research paper in 2005 found that an increase of collaboration has a significant correlation with decrease of variance of published quality (Rigby and Edler, 2005). Rigby and Edler conclude that collaboration acts as a "peer review" system and that "this peer review effect is inherent throughout the research process" (Rigby and Edler, 2005, 784).

The above mentioned studies offer narrow insights into specific management aspects in the context of medical research. The body of literature covering concrete management practices however, is limited to descriptive reports and case studies. For instance, a case report analyzing the very successful research lab of Professor Langer at Harvard University, points out that creating flat hierarchies between early PhD-students and postdoctoral candidates conveys the researcher a sense of responsibility, enables them to have successful knowledge exchange between members of the research lab and will create more successful researchers in the end (Bowen and Gino, 2006). Unfortunately, there is no data and no methodology to scientifically explore these claims. Conclusively, only isolated aspects of management such as collaboration patterns, incentives and prestige of the researcher have been analyzed. There nevertheless exists a clear lack of knowledge on the relationship between established management practices and research outcomes. This thesis is making efforts to create this knowledge.

THIS PAGE INTENTIONALLY LEFT BLANK

Chapter 3

Study Data And Methods

The data set is based on the medical research version of the World Management Survey (WMS). The methodology of the World Management survey has been previously thoroughly explained (Bloom and Van Reenen, 2007; Bloom et al., 2012b) and has been adapted for this new context of research (detailed description in Hillen (2016)). It employs an interview-based evaluation tool covering 18 management practices rated on a scale of one to five, which are categorized into four key dimensions of effective management practices: operations, performance, target and people management (see Appendix B). The adaption of the WMS to fit the medical research context was done using multiple expert interviews, iterations and internal validation tests. For this study, I identified principal investigators at Harvard Medical School who employ at least three full-time researchers within their laboratory. Together with a research team, I obtained data from 133 medical research laboratories at Harvard Medical School. The interviews lasted about 45-60 minutes and were conducted over the phone with a double-blind survey technique. A second scorer listened to the interview as well and debriefed with the interviewer afterwards to mutually agree on a score. In total, all interviews were split between a team of an interviewer and a listener, consisting of myself (Interviewer B) and three temporary research assistants (one interviewer and two listening scorers).

The interviews were administrated in September 2015 and from May-July 2016. The survey had a response rate of 33% and the respondents came from 18 different departments and 13 different institutions within Harvard Medical School. After checking for comparative criteria across the labs (e.g. eliminating very young or small labs), 117 research labs are analyzed for the results of this study. The respondents range from Assistant to Full Professors.

3.1 Survey explanation

The primary independent variable of this study is the overall management score from the WMS. It is the laboratory's average overall management score from one to five, with a score of five being the highest, across 18 questions. Besides keeping track of the demographics of a lab, the survey additionally covers information on five organizational structures, and eleven self-reported outcome variables.

3.1.1 Management practices

The main management score was computed as the average across the different management variables in four different domains of management: operational, monitoring, target, and people management. The management practices are based on the original WMS survey and most of the concepts are grounded in existing management literature. For instance, the dimension of operational management is based on the literature about dynamic capabilities (Eisenhardt and Martin, 2000) and organizational routines (Becker, 2004). Both concepts, dynamic capabilities – which describe the process of a company to flexibly adapt its resources to external changes – and organizational routines – describing the role of repetitive processes within e.g. corporations – inform the survey's measurements about organizational processes for effective management.The dimensions of monitoring and target management are to an extent founded upon the managerial framework of the resource-based view (Barney, 2001), describing how to obtain a competitive advantage through strategic resources. Accordingly, the survey is related to the idea that firm-specific assets and processes are closely related to sustaining a competitive advantage and a company's performance. Lastly, the people management practices are closely related to literature by Lepak et al. describing the importance of the human resource systems paying attention to employees' motivation, knowledge skills and opportunities to contribute (Lepak et al., 2006; Lengnick-Hall et al., 2009). Furthermore, the questions targeting the reward and promotion of employees, or here of the researcher, is grounded in previous literature around innovative work practices, such as incentives and training (Ichniowski et al., 1997). The following presents a short description of all eighteen management practices according to their managerial domain (see Appendix B for the detailed survey).

1. Operations

   Standardization and Protocols: tests standardizes main processes in the lab, such as experiments and operations.

   Rationale for standardized processes: tests motivation and impetus behind changes to operations and what change story was communicated.

   Continuous improvements: tests processes for and attitudes about continuous improvement and whether learning is captured/ documented.

   Good use of human resources: tests processes for and attitudes about collaboration and knowledge exchange between researchers.

2. Monitoring

   Performance Tracking: tests whether performance is tracked using meaningful metrics and with appropriate regularity.

   Performance Review: tests whether performance is reviewed with appropriate frequency and indicators.

   Performance Dialogue: tests the quality of review conversations.

   Consequence Management: tests whether differing levels of performance of projects (not personal but plan/ process based) lead to different consequences.

3. Target setting

   Types and Balance of Targets: tests whether targets cover a sufficiently broad set of metrics and whether quantitative and qualitative targets are balanced.

Interconnection of Targets: tests whether targets are tied to the organization's objectives and how well they cascade down the organization.

Time Horizon of Targets: tests whether the lab breaks down research questions into reasonable sub-experiments and has a short-, medium- and long-terms goals in planning and targets.

Target Stretch: tests whether targets are based on a solid rationale and are appropriately difficult to achieve.

Clarity and Comparability of Goals: tests how easily understandable performance measures are and whether performance is openly communicated to staff.

4. People management

   Rewarding high performers: tests whether there is a systematic approach to identifying good and bad performers and rewarding them proportionately.

   Removing Poor Performers: tests how well the organization is able to deal with underperformers.

   Promoting High-Performers: tests whether promotion is performance based and whether talent is developed within the organization.

   Managing Talent: tests what emphasis is put on overall talent management and continuous learning within the organization.

   Attracting Talent/Recruiting process: tests the strength of the employee value proposition.

3.1.2 Structures of the lab

As described in the literature review, most research labs have similar organizational structures, i.e. an existing affiliation with a university, and similar promotion and hiring systems. However, conducting extensive interviews reveals subtle differences in these structures. This study's sample is highly selected consisting only of HMS labs, still it is interesting to explore these variations across labs further. Given the importance of collaboration as seen in the literature review, I focus the first four questions on structures dealing with collaborative projects, tasks and targets. The fifth question is about the utilization of alumni, which represents an adaptation of the original WMS about "Retaining Talent". While in a corporate environment retaining talent is an important factor for human resource development (Bloom et al., 2007), as described in Section 2.1.1, in a medical research lab it is the objective of Human Resources not to retain, but to develop and send talent off to become independent researchers. Thus, I am interested in the impact of the alumni network on the long-run capabilities of a lab in terms of exposure, resources, collaborators, and other network effects. Lastly, testing the PI's frequency and attitude toward applying for research funding is grounded in the interest to better understand how the PI spends and prioritizes his time. For instance, it is estimated that especially new investigators have to resubmit 70% of their funding applications (Sapienza, 2004). This is quite significant, considering that each application takes a significant amount of time and attention of the PI away from focusing on research, managing his/her lab or mentoring his/her students. These assumptions led to the formulation of the following six organizational structures and respective answers.

1. Collaboration of projects: tests if the lab has one project per researcher, or multiple projects per researcher without and with collaboration.
2. Collaboration on manuscript writing: tests if the writing of a paper is done in a collaborative way (i.e. significant share of writing is done by both the PI and the lead author).
3. Decision power: tests who decides in the lab when and where to publish a paper, namely either the PI, the leading researcher, or both in collaboration.
4. Target setting: tests who sets the target and milestones for a research project, namely the research, the principle investigator or both.
5. Utilization of alumni: tests how much contact the PIs have with their alumni and to what extent they collaborate.
6. Applying for funding: tests when the PI applies for funding and if it is demand driven,

continuous, or opportunity driven.

As described in Section 2.1.2, there is no perfect output measure for medical research. To best approximate research outcomes and "performance" in terms of relevance and impact, I decided to collect various kinds of performance indicators. Part of the self-reported outcome variables can also be collected objectively via external sources, such as e.g. the publication via e.g. Pubmed or funding via the NIHReporter, which can be used to test the reliability of the interviewee. Furthermore, I ask for outcome variables which are difficult to collect as no exhaustive record exists online such as the career of alumni, applications received and conferences they went to. Collecting these variables (see below for an exhaustive list) helps to control for various aspect of the research lab and PI such as e.g. exposure (number of conferences) or attractiveness of the lab (number of applications). Also, these variables can be analyzed in terms of being outcome and success variables of good management in a research lab as well.

3.1.3 Self-reported outcomes

- Publications: number of publications the lab produces and number of papers the PI reviews each year.

- Funding: amount of funding per year, the success rate on funding applications, and what percentage of researchers in the lab have their own funding and how many grant applications the PI reviews per year.

- Other success criteria: percentage of alumni going to academia, number of patents, number of applications received and to how many conferences the PI goes per year.

3.2 Obtaining research surveys across HMS

This study defines an eligible lab as a research lab with (1) a principle investigator as the lead investigator and (2) by having three or more full time research scientists (e.g. postdoc, PhD student etc.) working in the lab. I reached out the professors of labs at Harvard Medical School in two batches. In the first, smaller batch, described in Hillen (2016), I had pre-screened research labs for these eligible criteria and afterwards, reached out to 100 eligible research labs. This approach resulted in 24 conducted interviews, four PIs who were willing to participate but only had time after my project timeline and 24 preferred not to participate in the research study, leaving 48 PIs non-responding. Overall, this first outreach had a response rate of 24%. In a second initiative, together with a small research team of three temporary research assistants, I had pre-screened 223 eligible research labs and then, in order to scale up the efforts, reached out the remaining Harvard faculty of 3138 principle investigators per email. In this email, I asked them if they meet the eligibility requirements and if they could quickly answer either "yes", "no", or "no - don't meet requirements" to indicate their eligibility and if they want to partake in the survey study. In order to obtain a high response rate, we carefully tracked the responses to the emails of the research labs, and followed-up after one week if the lab has not yet answered. For the pre-screened research labs, knowing that they were eligible, myself and one temporary research assistant also followed up with a call to the assistant to schedule an interview with the lab's PI. Overall, calling the PIs has shown to be most effective strategy to schedule an interview with a research lab. Out of the 223 eligible labs, 66 PIs participate in the interview and out of the 3138 Harvard faculty, 43 interviews were conducted. In order to estimate the response rate for the large batch of 3138 faculties, I screened a subsample of 985 of these research labs and identified 33 research labs as eligible. Taking this 3.4% (33/985) eligibility rate as a proxy for the entire set, I estimate that 107 research labs within the outreach of 3138 labs are meeting the eligibility requirements. Taking these numbers, this results in a response rate of 33% having conducted 110 interviews out of 330 (223 + 107) eligible PIs. In general, I obtained 150 positive responses (i.e. "Yes") of which

110 converted into an interview and 40 did not follow-up after their initial agreement. Of the 216 negative responses, 120 did not wanted to participate in the interview and 96 answered that they were not eligible. The remaining 2295 research labs did not reply. Overall, we had a response rate of 31% with having contacted 430 eligible PIs and conducted a total of 133 interviews.

3.3 Maximizing interview quality

When conducting the interviews, the interviewer and listener followed several steps to obtain high-quality responses. First, this study uses a double-blind interview technique. This method means that the team of interviewer and listener conducted the phone interviews without notifying the principle investigators that their answers would be scored on a scoring grid. This process helps to understand the actual management behaviour, instead of hearing about the theoretical best case scenario of the principle investigator. On the other hand, the interviewers and scorers were also given only the information necessary. They were only given the interviewee's name and telephone number but otherwise had no preconception about them nor were allowed to conduct any research on them beforehand. Second, the interviewers used open-ended questions avoiding leading the interviewee to a particular answers. For instance, in the second question of the dimension "performance monitoring", we ask "How do you review [previously mentioned indicators] for your lab?", instead of a close-ended question such as "Do you use a protocol to keep track of the project's progress ?". This first question is then followed, if applicable, by a second open question: "Can you tell me about a recent review meeting?" and then followed by "Who is involved in these meetings?". If necessary, the interviewers can ask further follow-up questions. The response to these questions are then scored against a scoring grid from one to five. One (1) is defined as 'Performance is reviewed infrequently or in an un-meaningful way (e.g. only the entire lab meets once in a while)', three (3) as 'Performance is reviewed periodically with both successes and failures identified; Results are communicated to everyone; There are lab meetings and “in-person” meetings' and five (5) as 'Performance is continually reviewed (including meetings and reports), based on different progress/success indicators; All aspects are followed up to ensure continuous improvement; Results are communicated to all staff'. Third, I conducted rigorous training with interviewer A and the two listener beforehand. Two full days, each interviewer and scorer had an introduction into the survey tool, followed by listening and double-scoring five previously recorded interviews. Furthermore, the first minimum five interviews were listened into by myself as well as I had weekly check-ins with the team and listened into several recordings to ensure a continuously high level of quality over time. Finally, the interviewer and listener collected noise-controls, specific variables evaluating the interview process itself. They kept track of the time spend on the interview, the willingness of the interviewer to share information, the patience of the interviewer as well as who was the interviewer and who was the listener. These noise controls are included in the regression analyze to improve the precision of my estimates and reduce the measurement errors.

3.4 Ex-post controlling for interview reliability

As described before, each interview was conducted by an interviewer, Interviewer A or B (myself), and a listener, Listener A or B. Both independently scored each management practice from 1-5 and thereafter agreed on one score. While I use the final, agreed score for all of the analyses in this paper, I can use the individual scores to test the robustness of the interview technique and management scores. As can be seen in Figure 3.1, of the 133 total interviews, 102 have been interviewed by interviewer A and 31 have been interviewed by interviewer B. Interviewer B, the author of this study, obtained extensive interview training and in a pilot

	Listener A	Listener B	Interviewer A	Interviewer B	
Interviewer A	57	42	-	3	102
Interviewer B	-	-	5	26	31
	57	42	8	31	133

Table 3.1: Interview count of interviewers and listeners

test of the survey described in Hillen (2016), 26 interviews were conducted without a listener second-scoring the interviews. As can be seen in Figure 3.1, the five (5) and three (3) interviews that were conducted with one of the interviewers listening into the interviews have been conducted for training purposes. In selected interviews, I listened as a third participant into the interview to ensure its quality; this is not represented in Figure 3.1.

3.4.1 Robustness check for Interviewer and Listener scores

The majority of interviews have been conducted by Interviewer A with either Listener A or B second-scoring the management practices. Thus, in the following robustness analysis, I focus on comparing the scores of Interviewer A and Listener A and B.

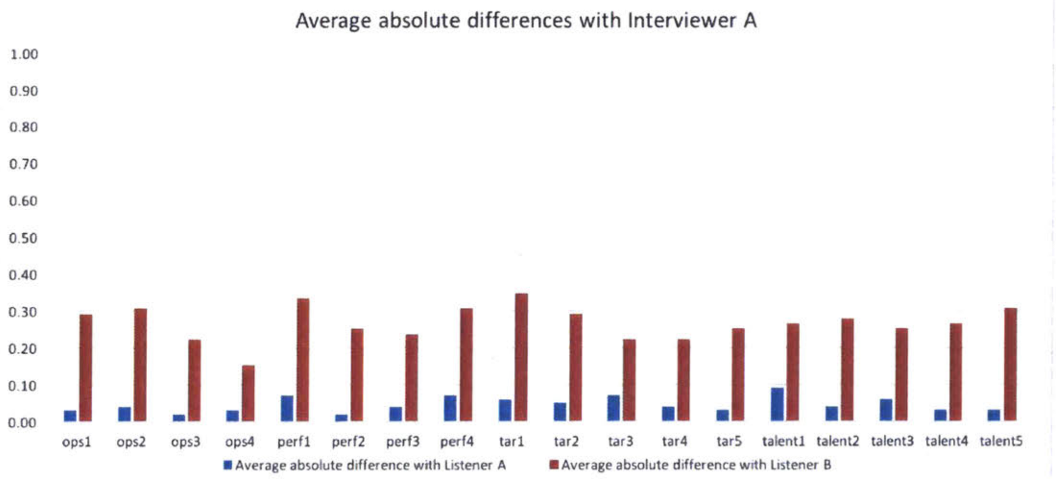

Figure 3-1: Average absolute differences with Interviewer A

The Figure 3-1 shows the average absolute difference of management scores between the interviewer and Listener A or Listener B across all 18 management practices. For instance, the blue bar for the variable org1 shows that the average absolute difference in scoring of Interviewer A and Listener A for the first management practice (i.e. operations management - question 1) across all 57 interviews. The difference in the score is almost nonexistent for Listener A. Also, for Listener B this difference is very small with an absolute difference significantly below 0.5

scoring points across all management practices. These results substantiate the robustness of the management scores for the 102 interviews conducted.

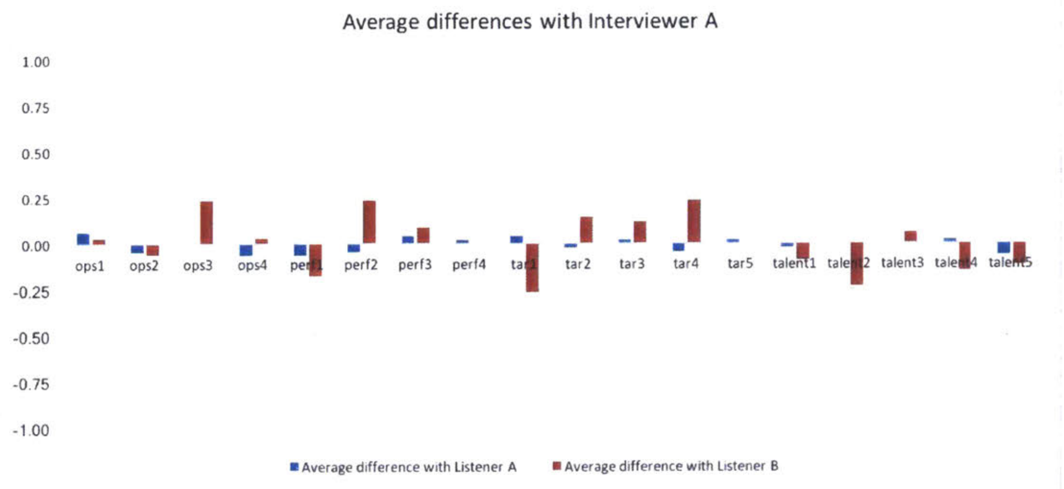

Figure 3-2: Average differences with Interviewer A

The Figure 3-2 shows the average difference of management scores, without taking the absolute value. Thus, I can analyze if the Listener has, on average, rated a particular management question as higher or lower than the interviewer. For both Listener A and Listener B, there seems to be no clear or consistent pattern of higher or lower ratings than those of Interviewer A. For example, Listener A rated 7 times higher and 8 times lower than the interviewer. Moreover, Listener B scored the labs 9 times higher and 7 times lower than the interviewer. As expected, drawing from Figure 3-1, Listener B has a slightly higher level of disagreement with Interviewer A than with Listener B.

The Figure 3-3 shows the correlation between the scores of Listeners A and B with the scores of Interviewer A. This is the pair-wise correlation between the Listener and the Interviewer score across all scored research labs, for each particular management practice. The second to last bar shows the average correlation across all management practices (i.e. the average of all previous bars) and the last bar shows the correlation of the final average lab scores between the interviewer and listener. Similar to the widely used Cronbach's alpha (Cronbach, 1951), the correlation between both independent scores should serve as a lowerbound reliability estimate

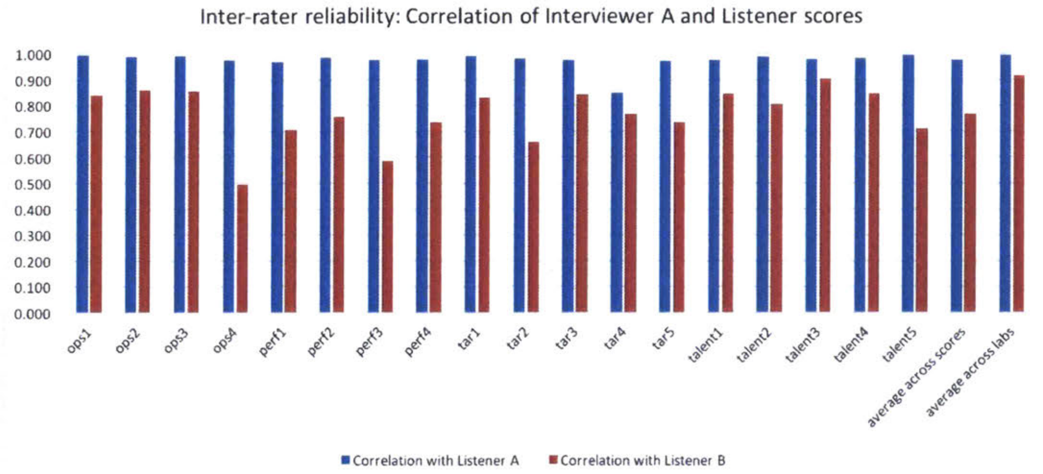

Figure 3-3: Inter-rater reliability: correlation of Interviewer A and listener scores

of the scores. If there is no disagreement or if there is a disagreement, but it is consistent (e.g. Listener A always scored 0.5 points lower than Interviewer B), the coefficient is 1.0. A negative correlation would, on the other hand, mean that both, interviewer and listener, have scores that move in opposite directions. While one scorer evaluated the lab to have good management practices, the other one judged them as having bad management. Orienting at the Cronbach's alpha again, I can use its previously defined rule for describing internal consistency (Kline, 2013) as a proxy for the reliability of my scores Table 3.2:

Cronbach's alpha	Corr. Coefficient	Internal consistency
$0.9 \leq \alpha$	0.9 ≤ Corr. Coef.	Excellent
$0.8 \leq \alpha < 0.9$	0.8 ≤ Corr. Coef. <0.9	Good
$0.7 \leq \alpha < 0.8$	0.7 ≤ Corr. Coef. <0.8	Acceptable
$0.6 \leq \alpha < 0.7$	0.6 ≤ Corr. Coef. <0.7	Questionable
$0.5 \leq \alpha < 0.6$	0.6 ≤ Corr. Coef. <0.6	Poor
$0.5 < \alpha$	0.5 <Corr. Coef.	Unacceptable

Table 3.2: Approximated rules for describing internal consistency

When analyzing Figure 3-3, it can be shown that Listener A is especially aligned and consistent with the scores of Interviewer A. In nearly all individual management practices the correlation coefficient is above 0.90, the average correlation coefficient across the scores is 0.97, and the average coefficient across the labs is 0.99. Thus, referring to Table 3.2, the internal reliability seems to be excellent between the scores of Interviewer A and Listener A and the agreed scores of both seem to be robust for the analysis. Analyzing the correlation of scores with Listener B, the disagreement is less consistent. Most management practices do have a higher correlation coefficient than 0.8, but there are some outliers, such as the fourth operation management practice having a correlation of less than 0.5. In general, however, the average correlation coefficient across all management practices is 0.76, which is arguably still very high. The average correlation coefficient between Interviewer A and Listener B is with 0.91, which is also very high as well and considered to be "excellent" (see Table 3.2). In most of my analysis in Chapter 5 I do use the overall management score and with 0.91, the overall management score is highly consistent and substantiates the robustness of the data.

3.5 Objective outcome variables

In order to understand the diverse outputs a research lab produces, I collected data from Scopus (Elsevier, 2017) consisting of summary statistics of all scientific articles. Specifically, I collected data about all published journal articles of the principal investigators such as title, journal, authors, citation count per year and in total as well as other contextual information. Furthermore, I obtained the funding received by the sampled labs from the National Institute of Health (NIH) for the last five years using the NIHReporter (NIHReporter, 2017). For the analysis, I calculated the funding the PI received by the NIH for the past five years before the interviews in which he was listed as the primary contact and first principle investigator. Usually, these grants are then used in his own research lab, while PIs being listed as Co-Investigators need to share grants with other research labs. Additional data was obtained from PubMed to classify research labs into different research fields. Over 210,000 medical subject headings (MeSH terms) from all publications of the sampled PIs were collected for clustering the labs into specific research fields (see Chapter 4).

THIS PAGE INTENTIONALLY LEFT BLANK

Chapter 4

Clustering

4.1 MeSH terms

This analysis explores the relationship between management and three outcome variables: number of publications, citations, and funding. It is important to note, however, that different fields within medical research vary significantly in terms of audience, reach and demand for management. Not controlling for the research field could negatively confound this study in two ways:

1. Some research fields, such as general biology, have a larger audience and a higher citation count which is not necessarily driven by better or worse management but by the field itself. Orthopedic surgery, for example, does not have general biology's typical counts.
2. Different research fields have different management demands and are thus differently impacted by management. For instance, while a biology lab conducting mice experiments might require significant managerial effort, a computational research lab handles analyses on existing data sets, and will therefore have very different management requirements.

Overall, not controlling for the actual research fields of the labs would bias the results in a way that I might not measure the direct association of management with publications, citations and funding but instead measure the effect of the research fields on either the management score or these outcome variables. An easy but imperfect option to control for the research fields would be to take associated institutions or the departments of the research labs as control variables. However, throughout this research it became apparent that while two research labs might be in the same department, it does not mean they are conducting comparable research in terms of management requirements and outcome variables. For instance, while several investigated research labs are in the "Genetics" department, it can mean everything from working with existing genomic data to conducting gene-editing experiments in wet labs. Thus, I need measurements that capture the actual research the sampled research lab is conducting. I achieve this by clustering the labs on the basis of their Medical Subject Headings (MeSH-terms).

Medical Subject Headings are hierarchically-organized terminologies describing, indexing and cataloging biomedical information (Lowe and Barnett, 1994). All MeSH terms, also called descriptors, form cumulatively the MeSH-tree with 16 main branches and numerous of sub-branches, leading to over 28,000 unique MeSH-terms as of May, 2018 (see an illustration of the tree in Figure 4-1). In order to illustrate the structure of this terminology tree better, consider the following example: The descriptor "Digestive System Neoplasms" has the tree-index C06.301. The first letter, C, is the first level of the tree, one of the 16 main branches, and stands for Disease. The first level, C06, stands for "Digestive System Diseases" and the second level, 301, for Neoplasms. Thus, with increasing depth of the tree, the terms become increasingly specific.

Every publication on PubMed is described by several of these MeSH terms. Hereby, the general guidelines prescribe that every concept which has been substantively discussed in the publication should be represented by a MeSH term. Thus, the number of MeSH terms can vary from publication to publication and the level of the term, the deeper the level the more specific the research concept or method is, can vary as well.

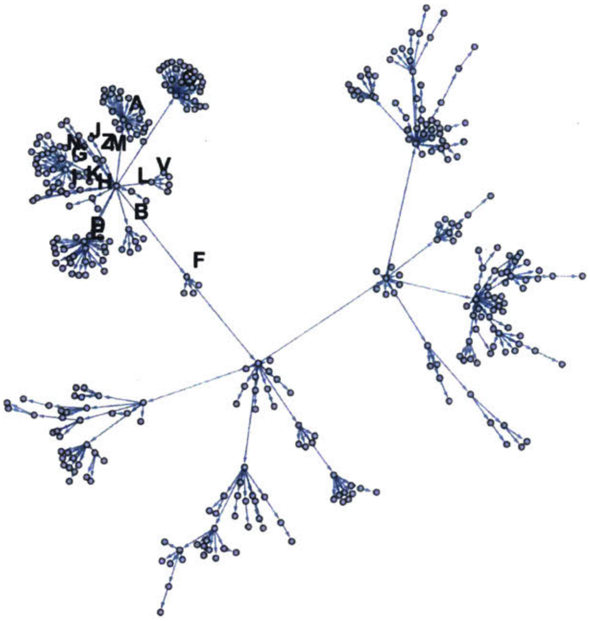

Figure 4-1: Visualization of the main branches of the MeSH tree
The 16 main branches of the MeSH tree and the first level of branch F (Psychiatry and Psychology) can be seen.

4.2 Data preparation

In order to obtain the MeSH terms for the publication of the observed research laboratories, I obtained all MeSH terms from all publications before May, 2016 from the 133 PIs using the PubMed API. As the PI is on all the publications published by his/her research lab and the, usually little, research he/she might conduct outside of his research lab will most likely be about the same research area as in the lab, I take the publications of the PI as a reasonable proxy for the general research area of the research lab. In total, this results in over 210,000 MeSH terms for 20,400 publications. The following analysis is restricted to the MeSH terms of the analyzed 117 research labs. As described before, the associated MeSH terms can occur at any level of the tree. To illustrate an example: One MeSH term of one publication of my sample can be Hemophilia A (tree-index: C15.378.100.100.500), an inherited genetic disorder preventing the body from creating blood clots which prevent bleeding, meaning that this publication most likely investigates some aspect of this disorder. In order to prepare the data for clustering, I go through all PIs and count the number of times each MeSH term is mentioned in their entire body of publications. This results in a matrix of 117 rows, each of them a PI representing a research lab, and the unique MeSH terms as columns with the counts of the terms as the values of the cells. However, using only the raw, assigned MeSH terms would have the result that two PIs who conduct very similar research might not be associated with each other. For instance, if one PI publishes many articles with the indexed descriptor "Blood Coagulation Disorders, Inherited" (tree-index: C15.378.100.100) and another one has many MeSH terms about "Blood Coagulation Disorders"(tree-index: C15.378.100), computationally they would have no overlap (see next chapter). Thus, I "back-propagated" each MeSH term and assigned a count of occurrence not only to the specific indexed MeSH term but also to each upstream term in the hierarchy. Following my example, I assigned a count to C15.378.100.100.500, C15.378.100.100, C15.378.100 up to the highest level of C (Disease). I further split the first level C15 into a first level C15 and the main branch C (Disease). The result is a matrix with 117 observation (the research laboratories) with 12,415 unique MeSHterms (columns). The resulting matrix is sparse with 106,310 cells containing a value with the overall sum of 306,821 MeSH term observations. As seen in Table 4.1, the deeper into the MeSH tree, the higher the count of MeSH terms for the 117 labs (i.e. a binary variable if a lab has a term at least once), the sum of the terms (i.e. summing up how often a term occurs). However, the dimensionality of the tree (i.e. the number of unique MeSH terms in my sample) also increases exponentially with increasing depth and thus, the density decreases significantly.

MeSH-tree level	**Unique MeSH terms**	**Count of terms**	**Density of resulting matrix**	**Sum of terms**
Main branch (0)	15	1572	0.90	58,723
1st	110	6,871	0.53	58,723
2nd	897	16,069	0.15	58,375
3rd	2323	26,043	0.0008	53,872
All	12415	106,310	0.0001	306,821

Table 4.1: Overview of the sampled MeSH tree across different levels

4.3 Clustering algorithm

Clustering seeks homogeneous subgroups across the observations in a multivariate data set. My goal is to partition the data into distinct groups so that the observations[1] within each group are similar to each other. I define similarity as having the same MeSH terms in their published work. In order to compute the distance between each observation which I use to estimate the similarity, I compute a distance matrix out of the matrix discussed before: Each row is an observation, i.e. PI, and each column an unique MeSH term. The value in the cells is the count of how often this MeSH term occurred in the publications of the PI. Each row then gets standardized by dividing each value by the sum of the row. I use the resulting matrix, X_{ij} to compute the Euclidean distance

$$d_{Euc}(x_i, x_k) = \sqrt{\sum_{j=1}^{p}(X_{ij} - X_{kj})^2} \tag{4.1}$$

between each observation, where x_i and x_k are the pairwise research labs which are compared and p is the length, i.e. the number of dimensions which are here the number of unique MeSH terms (125).

Having computed the distances between the observation, I group the data into clusters such that observations within each cluster are very similar (low distances) and observations in different clusters are dissimilar (high distances). For this task, I use K-means clustering, where each cluster is represented by the centroid (Euclidian mean of the data points) in each cluster. K-means clustering assigns the observation to predetermined number of clusters, K, in order to minimize the distances within and maximize the distances between the clusters. If C_i,....,C_K are the sets containing the indices of observations which are in the K clusters, then the within-

[1] the research labs represented by the PI

cluster variation with respect to the centroid positions is the following:

$$W(C_k) = \frac{1}{|C_k|} \sum_{i,i' \in_k} \sum_{j=1}^{p} (X_{ij}, X_{i'j})^2 \tag{4.2}$$

where X_{ij} and $X_{i'j}$ are the distance matrices of observation i' and i in cluster K calculating the distance across all dimensions p.

The goal is to minimize the following cost-function:

$$\sum_{k=1}^{K} W(C_k) \tag{4.3}$$

with respect to finding the best C_i,....,C_K.

The algorithm works as follows:

1. The algorithm initializes randomly K clusters, and assigns all observation to the nearest cluster centroid, thereby defining the first K clusters.
2. It repeats the following two steps until the cluster assignments converge, i.e. do not change anymore:
   (a) Compute the cluster cluster centroid $\overline{x}_k$ for each cluster k.
   (b) Reassign all observation to the k centroids and form new clusters based on the observations' closeness to the centroids.

4.3.1 Exploring the MeSH tree

Observing the increasing sparsity of the mesh-tree (see Table 4.1), I understand that there might be a trade-off between the extend to which I can differentiate between the labs and the depth of the MeSH tree, with the ability to create robust clusters for the 117 research labs. For instance, while the main branch might lead to robust clusters, the 15 different dimensions might not be sufficient to differentiate effectively between the labs. Thus, I first test different scenarios of clustering levels and the associated internal consistency of the resulting clusters. In order to measure the internal robustness of resulting clusters, I choose the average silhouette width. If a_i is the average dissimilarity, or distance, between observation i and the other points in the cluster to which i belongs and b_i the average dissimilarity between observation i and the other points in the next closest cluster to observation i. Then the silhouette for the observation i is:

$$s_i = \frac{b_i - a_i}{max(a_i, b_i)} \tag{4.4}$$

The average silhouette width is then defined as the average of s_i over i. It is a measure of how similar an object is to its own cluster (cohesion) compared to other clusters (separation). The silhouette ranges from 1 to +1 and the higher the value, the more it indicates that the object is well matched to its own cluster and poorly matched to neighboring clusters. In order to estimate how well my clustering algorithm works on the different levels of the tree, I compute the average silhouette width for two to six clusters. A high average silhouette width means the clustering configuration is appropriate. If many observation have a low or negative value, decreasing the average silhouette width, the clustering configuration may be not appropriate. As can be seen in Table 4.1, I calculated the average silhouette width for the main branch, the first, second and third level as well as all levels with the maximum depth of my sampled tree being the 12th level. I further investigated the cluster fit across the number of clusters from two to six clusters, calculating the overall average of the average silhouette widths to allow for quick comparison.

The first observation in Table 4.1 is that the isolated levels, first, second and third, do not result in good clusters from a metrical perspective. Except for number of clusters being two, they have a low average silhouette width between 0.09 and 0.19. Once, the levels are combined, the robustness of the clusters, the avg. silhouette width, increases. Thus taking isolated levels might not be the right approach as the clustering is missing a lot of contextual data, i.e. what is the broader research topic of this particular concept. In order to decide for the right level, I have to have a deeper look into the MeSH tree itself. The distribution of the number of MeSH

MeSH tree level	**Features**	**2**	**3**	**4**	**5**	**6**
Main branch (0)	15	0.46	0.25	0.28	0.28	0.28
First (1)	110	0.31	0.18	0.14	0.15	0.14
Second (2)	879	0.29	0.19	0.13	0.13	0.10
Third (3)	2323	0.17	0.18	0.09	0.09	0.09
0 & 1	125	0.39	0.18	0.20	0.20	0.20
0 & 1 & 2	1004	0.35	0.16	0.18	0.18	0.18
0 & 1 & 2 & 3	3327	0.34	0.15	0.16	0.16	0.16
All levels	12415	0.32	0.22	0.14	0.14	0.14

Table 4.2: Comparison of avg. silhouette width for different levels and clusters
This table shows a comparison of the avg. silhouette width for the isolated and connected levels one, two and three with the main branch and all levels across two to six numbers of clusters using K-means clustering.

terms for the main branch across the sampled 117 research labs is as follows:

- A. Anatomy (11%)
- B. Organisms (5%)
- C. Diseases (9%)
- D. Chemicals and Drugs (33%)
- E. Analytical, Diagnostic and Therapeutic Techniques and Equipment (16%)
- F. Psychiatry and Psychology (2%)
- G. Phenomena and Processes (18%)
- H. Disciplines and Occupations (1%)
- I. Anthropology, Education, Sociology and Social Phenomena (1%)
- J. Technology, Industry, Agriculture (∼0%)
- K. Humanities (∼0%)
- L. Information Science (1%)
- M. Named Groups (2%)
- N. Health Care (1%)
- V. Publication Characteristics (0%)
- Z. Geographicals (1%)

Over two thirds of all sampled terms are in the three fields, "Chemical and Drugs" (D), "Phenomena and Processes (G) and "Analytical, Diagnostic..." (E). Thus, clustering only on basis of the first level will likely not capture enough differences between the labs. Using all levels together would theoretically help better differentiate between the labs effectively. However, I only have 117 observations and their characterization in terms of MeSH-terms and depth differs significantly. The caveat is that going deeper into the tree introduces a bias which is based on the different depth of my sampled labs: If one lab has a publication which is described with the term "Bronchiolitis Obliterans" (C08.127.446.135.140) and another lab has a publication in "Bronchial Diseases" (C08.127), the back-propagation helps to match the two nevertheless as the first lab also receives an assigned value at C08 and C08.127. The issue though is that for the index-variables on the 5th, 4th and 3rd level (i.e. .140, .135, and .446), these two labs would be calculated as being unproportionally far from each other. This is because the importance of distances is not represented in my distance matrix. This means that the dissimilarity between a publication associated with the MeSH term A01 and another publication with the term Z07, is mathematically the same distance than between C08.127.446.135.140 and C08.127.446.135 from my previous example. While the latter of course has slightly different research as one is more specific than the other, they should be intuitively more closely associated to each other than a paper about B, Anatomy vs. L., Information Science. This bias seems to become more significant with increasing depth of the tree, thus I call it "depth-bias". This can be further illustrated by Figure 4-2. It shows the distribution of MeSH terms across research labs. On the x-axis are the unique MeSH terms represented in my sample for the first, second and third level. On the y-axis is the percentage of research labs, having at least one time an occurrence of this MeSH term. On the first level 50% of the labs cover around 50% of the MeSH terms. However, in the second and the third level already, only a small percentage of MeSH terms are covered by half of the research labs (less than 10%). This effect increases with tree depth, and the distribution of MeSH terms becomes wider going deeper into the tree.

Increasing depth will therefore increase this bias of overestimated distance between research

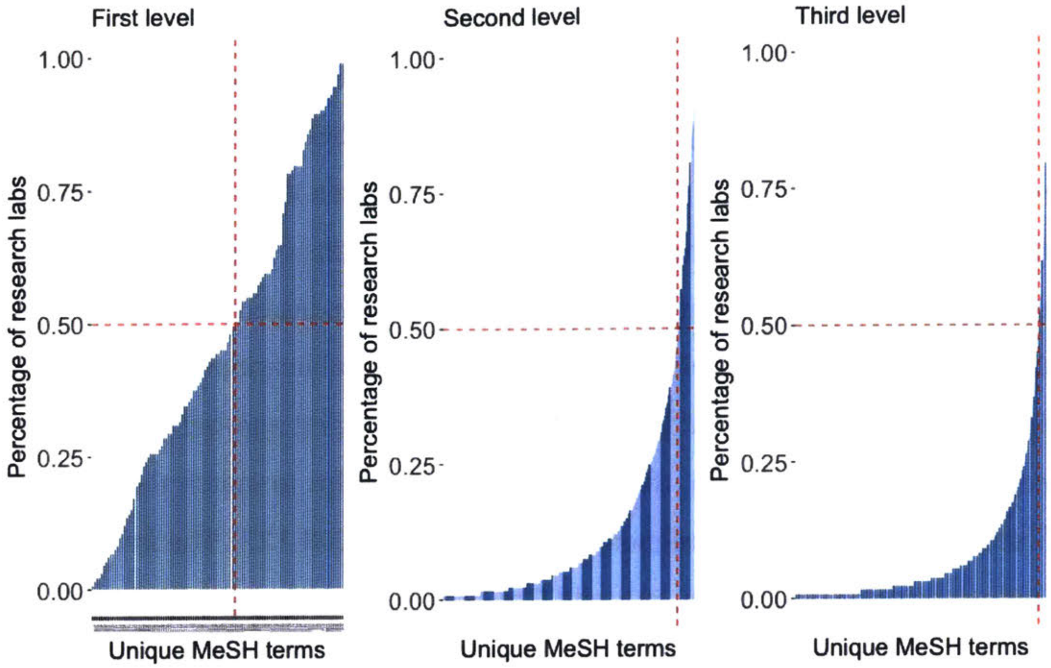

Figure 4-2: Distribution of MeSH-terms across levels one to three
The plot shows the distribution of MeSH-terms across the first (left plot), second and third level (right plot). On the x-axis are the unique MeSH-terms in my sampled research labs at the respective level and on the y-axis, the percentage of research labs having at least one occurrence of these terms.

labs. There are three factors to be considered when choosing the final level of the tree to compute my clusters on. First, I want to minimize the depth-bias just explained, second I do aim to get a reasonable internal robustness of the clusters and third I need to apply my own knowledge of the field to understand on what level of detail I differentiate best between the labs. First, combining the main branch and first level, seems to still yields an acceptable average silhouette width (Table 4.2). Second, on this level the depth-bias is not significant as most labs cover a significant share of sampled MeSH terms (see Figure 4-2). Third, making sense of the different levels in the original MeSH tree, shows that already the first level differentiates very well between the different topics. For instance, the first level aggregates specific research in the term Neoplasms (C04) which already indicates research on a very specific disease, which probably only a handful of the labs do. Going down one more level, instead of clustering on the basis of Neoplasms, I would differentiate between 15 different Neoplasms diseases (C04.588 (Neoplasms by Site) to C04.700 (Neoplastic Syndromes, Hereditary). The level of detail of terms seems to be unreasonable considering my small sample size of 117 observations. If further research has a much larger sample sizes or can integrate the distances in the tree into the distance matrix, it will make sense to go deeper into the MeSH tree to pick up stronger differentiation. For my research, taking the three points into consideration, I decide to cluster on the basis of the main branch and the first level (in Table 4.1 it would be the "0 1" row).

4.3.2 Finding number of clusters

For my K-means clustering algorithm I need to define K, the number of clusters, beforehand as an input parameter. In order to find the optimal number of clusters I use three different comparison metrics: Elbow-test, the silhouette plot and the Gap-statistic. The most straightforward and often employed idea of the elbow-test is to take the number of clusters K, which significantly decreases the within-cluster sum of squared distance, by visual inspection. Let W(C_k) be the within-cluster sum of squared distances between all pairs within the cluster k for one particular cluster, then

$$T_K = \sum_{k=1}^{K} W(C_k) \tag{4.5}$$

is the total within-cluster variation for the clustering with K clusters. I let K vary from one to ten and compute the T_K for each K. Plotting T_K against K, I can see a clear bend ("knee") in the graph (see Figure 4-3). Visually, the most clear "knee" is at K=2, another one might be at K=2. Often the "elbow" cannot be unambiguously identified and also here it seems that there is still a strong decrease of the within-sum of square after K=2.

Thus, a different, less subjective and more objective method, is the the average silhouette method. As described earlier in the chapter, I calculate the silhouette width with a high level signaling a good cluster fit. I calculate the silhouette width for one to ten clusters and pick the highest average silhouette width (see Figure 4-4). This would yield 2 clusters, meaning that the internal consistency would be strongest in the case of two clusters. Another small kink can

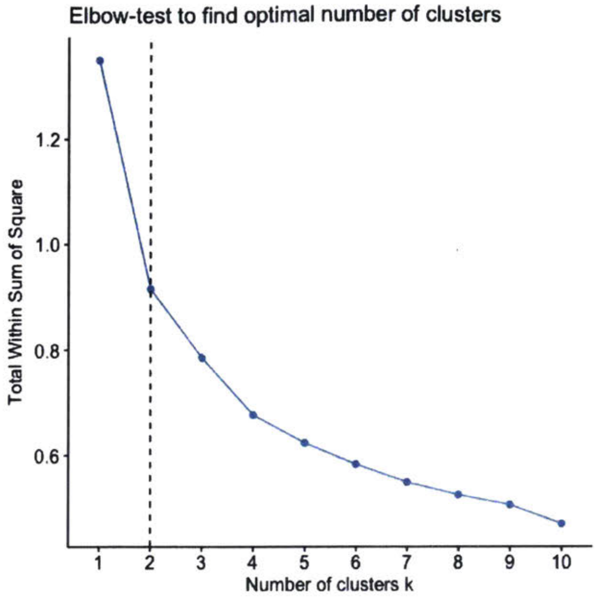

Figure 4-3: Total within-cluster variation for the clustering with one to ten clusters

be observed at K=5.

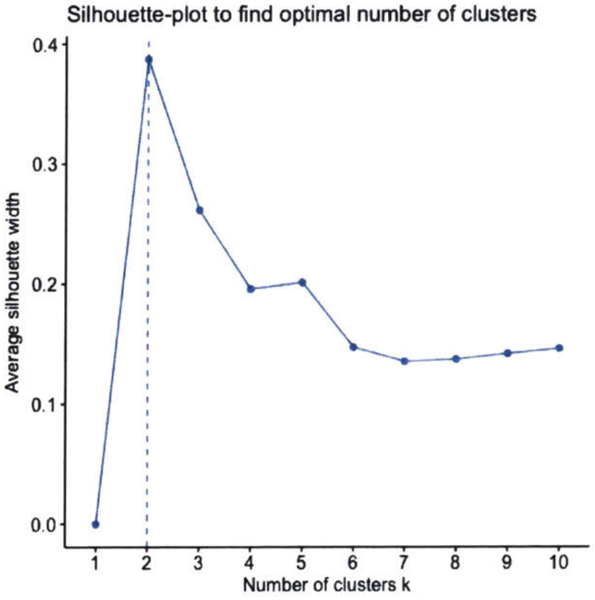

Figure 4-4: Total silhouette width for the clustering with one to ten clusters

The third method I want to use is the gap statistic. In this method, given a particular choice of K clusters, I compare the total within cluster variation to the expected within-cluster variation if the data would be randomly distributed (i.e., it would have no obvious clusters). The algorithm works as follows: I cluster the data for a varying number of clusters K from one to ten. T_K is the total within-cluster sum of squared distances. Next, one generates B random data sets of size n, with the simulated value of j uniformly generated over the range of the observer variable x_j. For each generated data set, B, one performs clustering for each K, from one to ten, and computes the total within-cluster sum of squared distances T_K^b. This is then used to calculate the Gap statistic:

$$Gap(K) = \left(\frac{1}{B} \sum_{b=1}^{B} log(T_K^b) \right) - log(T_K) \tag{4.6}$$

I can further compute the standard deviation with

$$\overline{w} = \frac{1}{B} \sum_{b=1}^{B} log(T_K^b)) - log(T_K) \quad \text{in} \quad sd(K) = \sqrt{\frac{1}{B} \sum_{b=1}^{B} \left(log(T_K^b) - \overline{w} \right)^2} \tag{4.7}$$

$$\text{resulting in} \quad s_K = sd(K)\sqrt{1 + 1/B} \tag{4.8}$$

Using this, I choose the optimal number of clusters using the metric introduced by Tibshirani et al Tibshirani et al. (2001) as follows:

$$Gap(K) \geq Gap(K+1) - s_{K+1} \tag{4.9}$$

meaning to choose the first K, if the Gap statistic of the consecutive K minus it's standard deviation is not exceeding the current Gap statistic (illustrated in Figure 4-5).I run this algorithm and set B to 500. This results in the gap-statistic plot in Figure 4-5. The result is that five clusters would give the best clusters according to the gap statistic, while the elbow method and the silhouette plot suggest two clusters as optimal.

Neither metric is perfect and it depends on my objective, optimizing for the internally most

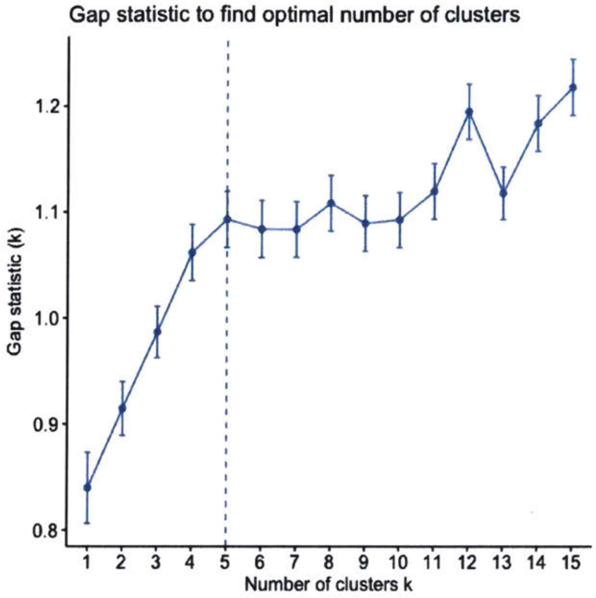

Figure 4-5: Total Gap statistic for the clustering with one to ten clusters

robust clusters, I would choose two clusters. However, going back to my original purpose of using clustering, I want to group research labs together to identify labs which do very similar research in terms of audience they publish to and management requirements they need. Overall my sampled labs comes from 21 different departments, ranging from clinical research at MGH to genetics at Dana-Farber Cancer Institute. Thus, it seems more sensible to have five over two clusters, as two clusters would not adequately capture the heterogeneity of my research labs. Thus, I proceed for the following clustering with K=5 clusters on the basis of the MeSH-tree from the main branch and the first level.

4.4 Resulting research areas

K-means offers a local optima but not necessarily a global optima. With each iteration, the starting points of K-means are different, thus every time I run K-means with K=5, I get slightly different results. In order to get the best result possible, in terms of minimizing $W(C_k)$, I run my k-means 1000 times and pick the smallest $W(C_k)$ (maximum of iterations until convergence=1000). The resulting clusters with assigned research labs are listed in Table 4.3.

Cluster	Research labs	Size
1	I04, I11, I30, I37,I40, I48, I57, I58, I87, I89	10
2	I01, I03, I06, I10, I103, I109, I113, I116, I12, I15, I16, I17, I20, I22, I25, I28, I31, I32, I35, I38, I39, I47, I49, I52, I56, I64, I66, I73, I74, I75, I76, I77, I83, I84, I99, I210, I211, I232, I237, I238, I239, I256, I278, I280, I295	45
3	I02, I09, I114, I13, I18, I34, I46, I51, I60, I62, I65, I67, I71, I82, I85, I95, I219, I231, I240, I247, I254, I270, I282, I284, I290	24
4	I07, I102, I108, I110, I111, I14, I29, I33, I36, I41, I44, I59, I68, I70, I81, I88, I90, I93, I97, I228, I229, I230, I267, I293	25
5	I05, I100, I101, I106, I112, I21, I24, I54, I55, I79, I80, I86, I94	13

Table 4.3: Final clusters and assigned research labs

They can be visualized by projected onto the first two principle components (Figure 4-6). Each ID associated with each point belongs to one research lab. The first principle component captures 85.7% of variance and the second 5.5% of variance. Cluster five seems to be quiet distinct from the other clusters. Especially cluster one and two seem to be quiet overlapping. However, this is only a visual representation of the first principle components and are not conclusively informative of how well the labs are clustered.

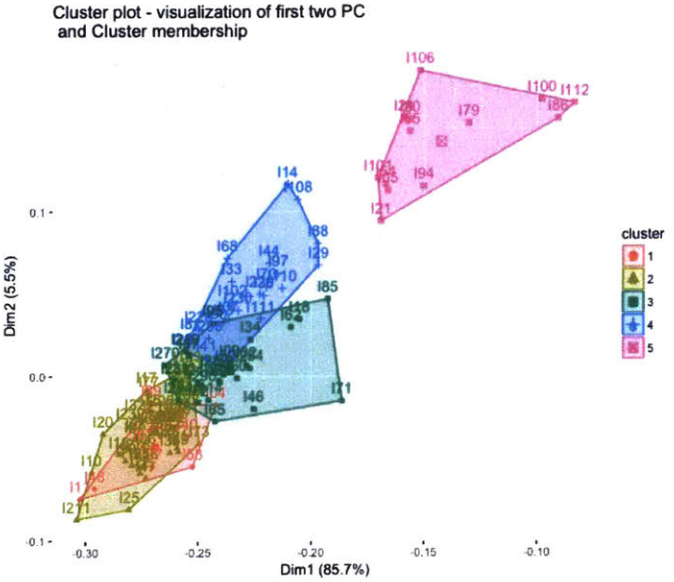

Figure 4-6: Visualization of clustering results and first two principal component

Another method of clustering, I use to cross-check my results is hierarchical clustering. It does not require choosing a number of clusters beforehand. I use agglomerative clustering, a bottom-up technique, in which each observation starts in its own cluster and at each iteration of the algorithm, two clusters which are most similar, given a dissimilarity measure, are combined into one cluster. This is repeated until all observation belong to one large cluster. As a dissimilarity metric I use the average linkage method, which computes all pairwise dissimilarities between observation in a cluster and calculates the average between them. The results can be visualized in a dendogram (see Figure 4-7). Hierarchical clustering seems to give slightly different results. It yields two small groups, one only consisting of one research lab, an outlier hard to be classified. However, when comparing the other groups there seemed to be a good overlap between the groups of the K-means clustering and the hierarchical.

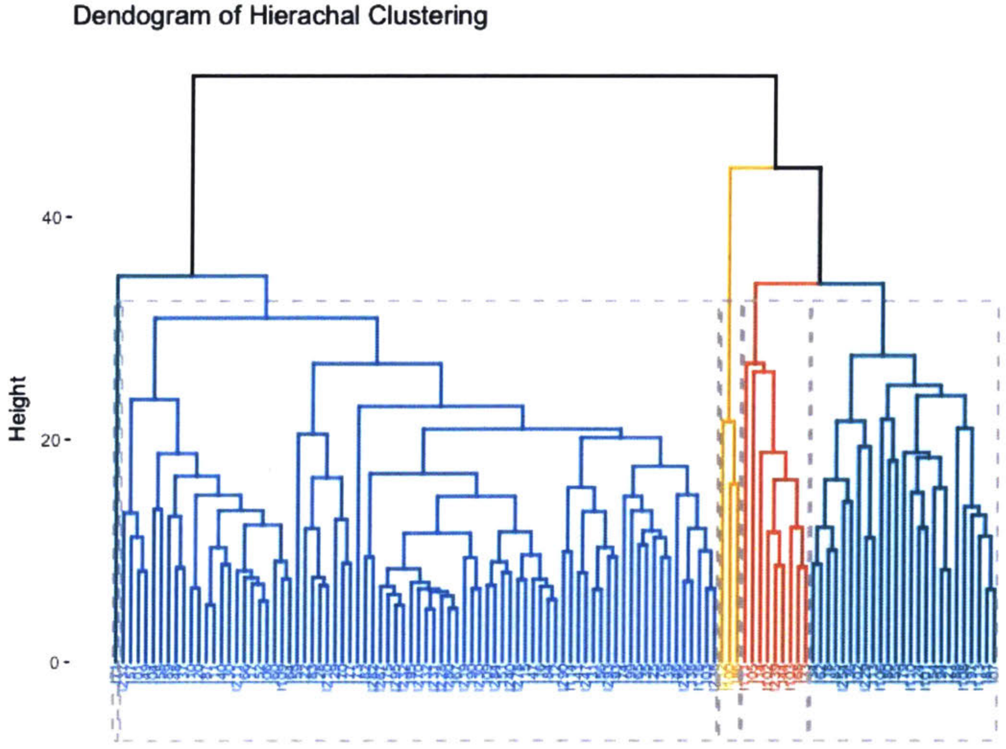

Figure 4-7: Dendogram of research labs and their respective groups

4.5 Ex-post sanity check

In order to control the performance of the clustering, I use two metrics. First, I use the ratio of within-cluster similarity to between-cluster similarity, with a higher ratio describing a better cluster solution (Kassab and Lamirel, 2008). Second, I look ex-post at the assigned research labs to check if the results are reasonable, i.e. if the research labs within a cluster have common characteristics. For the first check, I obtain the silhouette plot. As explained in the previous section, the silhouette width is a similarity measure. A high value of s_i 1, means that the observation fits very well into the cluster, a s_i 0 means the observation is between two clusters and a $s_i < 0$ means that the observation is probably in the wrong cluster. The Figure 4-8 shows that clusters have a clear positive value. Especially, cluster two and four seem to fit well. In total there are seven research labs which might belong in a different cluster, i.e. have a small negative silhouette width. The overall average silhouette width is 0.2. While it is hard to determine how well the clustering is on the basis of one number, this indicates that there are clear clusters but the internal robustness of them might be weak. This could be due to my little data of only 117 research labs and comparable many dimensions of 125.

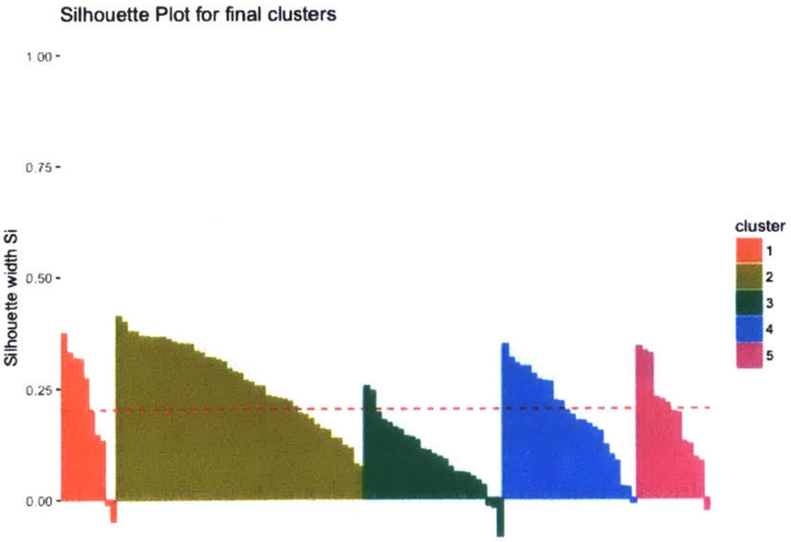

Figure 4-8: Silhouette plot for K-means clustering

I further want to understand what these cluster represent. Thus, I go through all 117 research labs' websites to understand what kind of research they are doing to understand commonalities of the clustered research labs. This assignment is, however, subjective: Some research labs do not write much or only on a high level about their work and most websites are focused on the research area but omit the actual methodologies used. This is the main reason I have to rely on this clustering instead of using my own guesses of research fields. However, when now trying to make sense of the clusters it is a useful exercise.

First, I only grouped the research labs into clinical versus non-clinical research. I define clinical research as either working with human subjects and/or in a hospital setting with the clear objective on having an impact on current practices. Furthermore, I plotted the five clusters against the clinical (red) versus non-clinical (blue) research estimate (see Figure 4-9). Interestingly, cluster five has almost only clinical research labs and the fourth has a large majority of labs conducting clinical research. The others do not have any clinical research. This gives confidence that my algorithm picks up the correct type of differentiation.

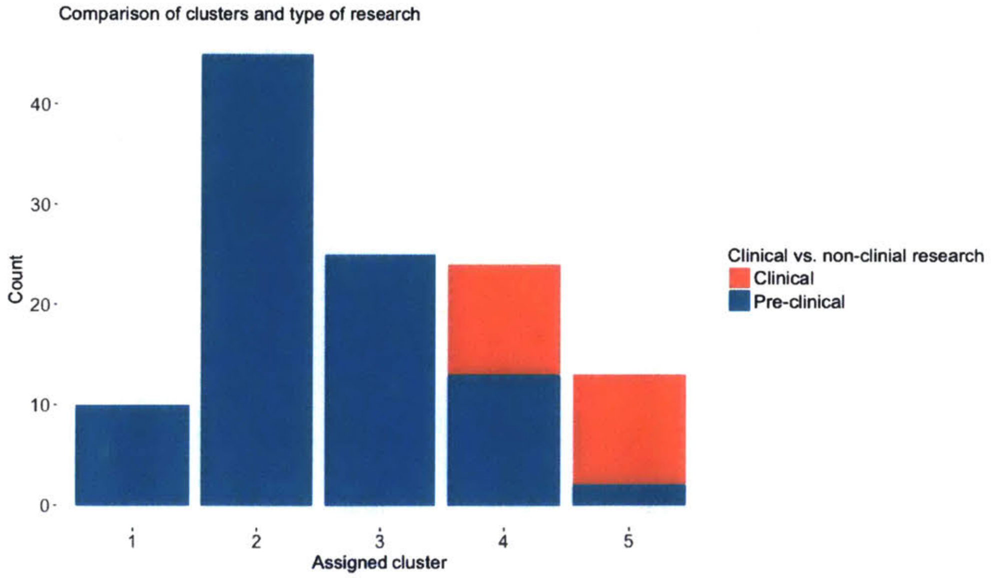

Figure 4-9: Final clusters compared with clinical and pre-clinical research classification

In a second step, I assigned an approximated research area to each of the research labs ranging from Biology to Surgery using their website's information to compare it with the resulting clusters. As can be seen in Figure 4-10, the research areas are distributed across all different clusters. Given that each research area varies in size this is not surprising. Nevertheless, there are some general patterns I can observe. The first cluster, seems to consist of biology, genetics, neurology and systems biology. The second cluster, is dominated by biology with almost half of the labs in this field, and the rest of the fields being in closely associated fields such as Immunology and Neurobiology. The third cluster, is strongly represented by Genetics and Pediatrics. The fourth cluster seems to be very interdisciplinary. It has many different research fields in it, especially research fields which have not occured anywhere else such as Dermatology, Surgery and Interdisciplinary research. This seems to be consistent with Figure 4-9, which shows a half/half split between pre - and clinical research. The last cluster, is strongly dominated by Radiology and Psychiatry. Both fields are in research heavily intervened and thus this gives me additional proof for the reasonableness of my clusters.

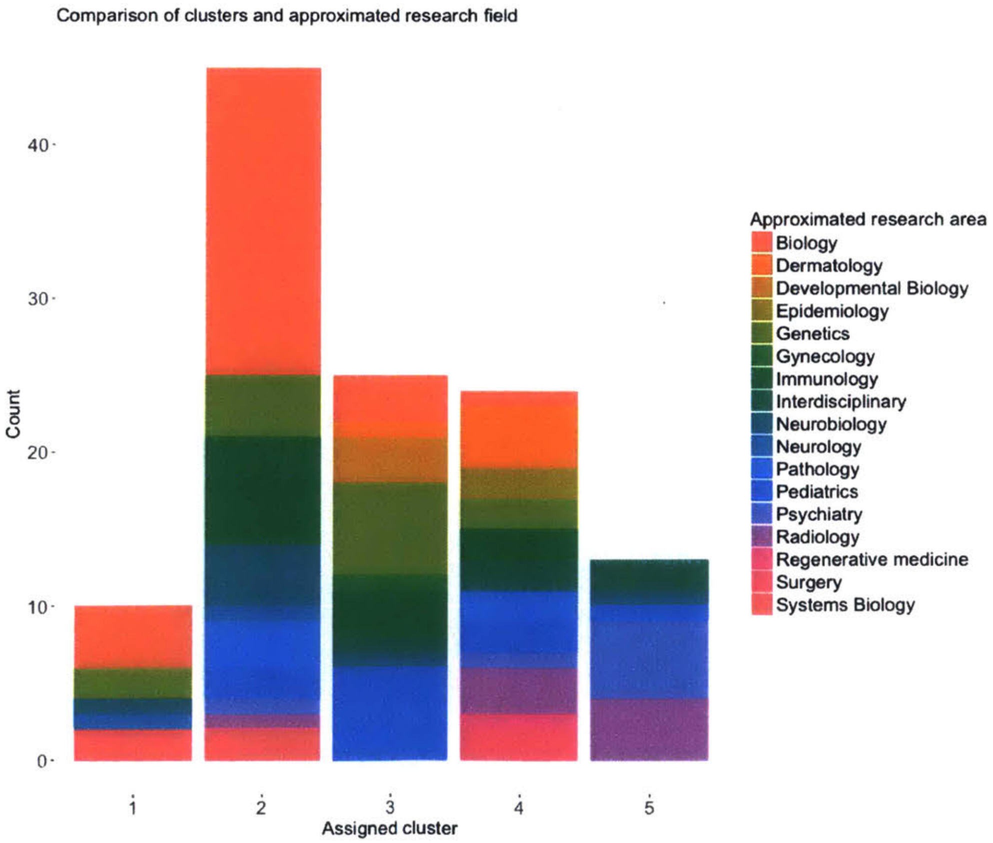

Figure 4-10: Final clusters compared with approximated research fields

Each of the five clusters contains research which is quiet similar in terms of their publications and with that the methods they use, the audience they address and the research they conduct. Using these clusters as control variables helps me in further analysis to effectively control for biases which are associated with research fields. For the purpose of interpretability I ex-post label these five clusters and assign them a rough research area. However, this is under the disclaimer that labelling these clusters definitely is not possible and these labels are the best effort possible to give them a common name. Looking at the different research areas and the analysis above I decide to label the five clusters as in Table 4.4.

Cluster	Label/Approximated research area
1	Neuro -& Systems Biology
2	Biology
3	Genetics
4	Interdisciplinary research
5	Psychiatry & Radiology

Table 4.4: Final clusters with assigned research area labels

Chapter 5

Statistical Analysis and Results

In order to answer the research questions of this study, I performs the following statistical analysis on the data discussed in the previous chapter. Consecutively, I present the results of the different parts of this study.

5.1 Statistical analysis

Before conducting an in-depth analysis, this study presents summary statistics of the demographics, overall management score and research output metrics of the sampled research labs. For the purpose of the entire statistical analysis, this study excludes the first two operational management practices, question one and two of the survey. The internal validation testing, i.e. comparing the cognitive understanding of the questions/formulations and their respective responses across all interviewees, found that there was no common understanding/interpretation of the questions across all interviewees. Their responses and scores seemed to be more associated with their understanding of the vocabulary than with the actual practices. This test however, gives further confidence in the reliability of the remaining management practices. With the cleaned data, I perform the following analyses.

First, I use descriptive statistics to gain a better understanding of the nature of the management score and the obtained data set. I examine the distribution of the management score, the differences of management scores across the four management domains and between- and within-variance across institutions and departments.

Second, I use multiple regression models to assess the relationship between the z-score management index for each domain of management (operations, performance, target and people) and the different laboratory's characteristics, such as lab age, number of researcher (log), PI's gender and research field of the laboratory obtained by my cluster analysis (see Chapter 4). The z-score in management is obtained by taking the average score across the management practices 3 to 18 and standardizing it by subtracting the average and dividing by the standard deviation of scores. It enables easier interpretation as an increase of one z-score management translates into an increase of one standard deviation in average management score within the sample.

Third, I run a multiple regression model with the three main research output variables (explained in Sections 2.1.2 and 3.4) publications, citations and funding as my dependent variables and the overall management score as my main independent variable. The number of publications are standardized by measuring publications per researcher and year with the PI as last author within the last five years before the interview. The citations are measured per publication in the last five years with the PI being last author and the funding represents the average NIH funding per year for the last five years with the PI of a lab as being listed as the primary contact. My model relies on robust standard errors and includes lab characteristics, such as lab age, number of researchers (log), PI's gender and research field of the laboratory, as well as a set of survey variables including interview duration, reliability and interviewer dummy variable. In order to check for the robustness of my analysis, I conduct this analysis with three sets of approximations of the "research area" of a lab: the research areas identified by my cluster analysis in Chapter 4, the institutions of the research labs and their departmental affiliations.

Fourth, I split the data set at the median into young (<15 years) and old labs (≥ 15 years) and conduct the same regressions as before, only differentiating between the labs' age to better understand the underlying drivers of the management impact.

Fifth, I categorized the laboratories into two groups: those with and without collaboration in research projects. A lab without collaboration is defined as having no active projects with more than one research being responsible for its success. On the contrary, a lab with collaboration were labs which had intentionally more than one researcher assigned to a research project.

5.2 Descriptive statistics

5.2.1 Summary statistics

Table 5.1 shows the summary statistics of the sampled research labs. The lab size, defined by the number of researchers, is on average 11.18 full time researcher (standard deviation of ±10.85) and lab age, defined as the time since the establishment of the research lab, is on average 14.83 years (standard deviation of ±8.71). Less than half of the research labs have a lab manager (44%). In this context, lab managers are responsible for running their laboratories safely and efficiently, taking responsibility for equipment, supplies and documentation. The control variables, the willingness to reveal information and the impatience of the interviewee, have both a very high average (4.80 and 4.57, respectively), reflecting the good will and interest of the interviewees in this study. The last three variables in Table 5.1 are objectively measured outcome variables. On average, a lab published 0.47 (standard deviation of ±0.33) journal articles per year per researcher and received 18.21 (standard deviation of ±13.48) citations per publication in the last five years. Since not all laboratories are funded by the NIH or have received their own research grant, only 108 labs occurred in the NIH data set. On average, these labs have $178,884 per researcher and year (standard deviation of ±$430,004) at their disposal with a notable variation across research labs.

	(1)	(2)	(3)	(4)	(5)
VARIABLES	N	mean	sd	min	max
Management score	117	3.122	0.477	1.806	4.333
Number of researcher	117	11.18	10.85	3	80
Lab age	117	14.83	8.710	2.200	44
Lab manager (0/1)	117	0.444	0.499	0	1
Willingness to reveal information	117	4.803	0.513	3	5
Impatience	117	4.573	0.813	2	5
Publication per year (PI last author)	117	4.231	3.233	0.200	17.20
Citations per paper (PI last author)	117	18.21	13.48	0	78.67
Funding per year/researcher	108	178,884	430,004	2,476	3.599e+06

Table 5.1: Summary statistics
The average management score is the average score across all 16 management practices ranging from one (lowest score) to five (highest score). Lab age is the time since the foundation of the research lab, willingness to reveal information and impatience are control variables collected in the survey data.

5.2.2 Management scores

The analyzed sample consists of 117 medical research labs at the Harvard Medical School. On a scale from one to five, the average overall management score for the sample is 3.12 (standard deviation of ±0.48). The distribution of the management score can be seen in Figure 5-1.

An overview about the summary statistics of the different management dimensions can be seen in Table 5.4. By comparing the summary statistics of the management dimensions with each other, I can shed light on how the dimensions differ in medical research. For instance, the mean of performance management (3.297) is significantly higher than of target management (3.038) (t-test is significant at the 1% level). Furthermore, by comparing the different percentiles (e.g. the difference of p75 and p25) I can better understand the distributions of

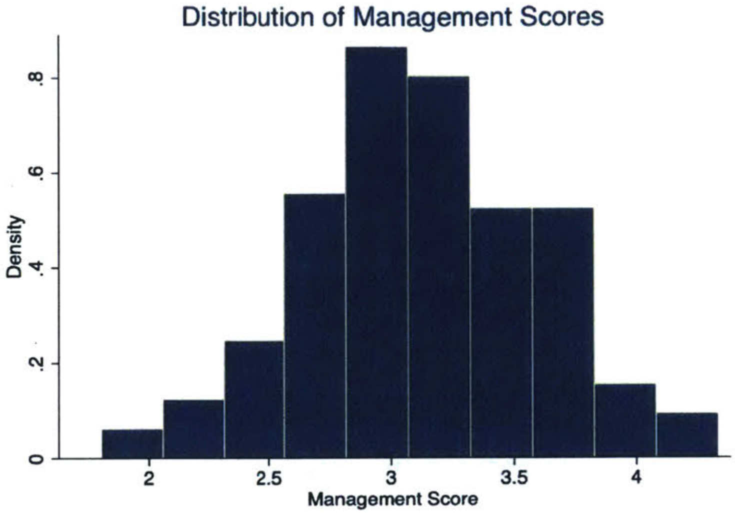

Figure 5-1: Distribution of management scores
Depicted is the distribution of the average management score from one to five, with a score of five being the highest, across sixteen management practices across 117 labs

the different dimensions. In operational management for example, the majority (50%) of research labs are narrowly distributed between a score of 3.0 and 3.5. For people management 50% of research labs have a score between 2.7 and 3.5 (see Table 5.4) .

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
VARIABLES	N	mean	sd	min	p25	p50	p75	max
Management score	117	3.122	0.477	1.806	2.833	3.111	3.444	4.333
Operational management	117	3.184	0.641	1.500	3.000	3.250	3.500	4.750
Performance management	117	3.297	0.692	1.750	2.875	3.250	3.625	4.500
Target management	117	3.038	0.504	1.700	2.800	3.000	3.400	4.300
People management	117	3.123	0.554	1.800	2.700	3.200	3.500	4.500

Table 5.2: Summary statistics for the different management dimensions
The management score is the average score across all 16 management practices. Operational management is the average score across the questions one and two of the survey, performance management represents the average of questions 5 to 8, target management of questions 9 to 13 and people management of questions 14 to 18.

5.2.3 Variation within and across institutions and areas

Furthermore, I analyze the variation of management score across and within institutions and research areas. As can be seen in Table 5.3, the standard deviation is significantly higher within the institutions (0.460 St.Dev.) than between the institutions (0.284 St.Dev.). The sampled research labs come from 13 different institutions, but the majority of labs is associated with the five teaching hospitals of the Harvard Medical School. Further, I compare the management variation across and within the research fields identified in the clustering ranging from Biology to Psychiatry (see Chapter 4) on basis of the actual publications and MeSH terms. As can be seen in Table 5.4, the within variation has a similar high standard deviation (0.469 St.Dev.) as the actual standard deviation of the management score, while there is only a low between standard deviation (0.097 St. Dev.).

Institutional comparison	(1)	(2)	(3)	(4)	(5)
Variables	Mean	Std. Dev.	min	max	Observation
Management score	3.121711	.4768748	1.805556	4.333333	N = 117
Between variation		.2839275	2.574074	3.638889	n = 13
Within variation		.460635	1.934211	4.290829	T-bar = 9

Table 5.3: Between and within institutional variation of management scores
In total the sample consists of 13 different institutions: Beth Israel Deaconess Medical Center (2), Boston Children's Hospital (13), Brigham and Women's Hospital (16), Dana-Farber Cancer Institute (15), Faculty of Arts Sciences(1), Harvard Medical School (23), Harvard School of Dental Medicine (3), Harvard School of Public Health (4), Immune Disease Institute (1), Joslin Diabetes Center (3), Massachusetts General Hospital (34), McLean Hospital (1), Veterans Affairs Boston Healthcare System (1). I compute the within and between variation of the overall management scores of these institutions.

Research area comparison	(1)	(2)	(3)	(4)	(5)
Variables	Mean	Std. Dev.	min	max	Observation
Management score	3.121711	.4768748	1.805556	4.333333	N = 117
Between variation		.0974308	3.005787	3.246667	n = 5
Within variation		.4691425	1.787361	4.315139	T-bar = 23.4

Table 5.4: Between and within research area variation of management scores
The research areas are defined using k-means clustering (see Appendix) and can be described as Systems Biology, Biology, Genetics, Interdisciplinary research and Psychiatry and Radiology. I compute the within and between variation of the overall management scores of these research areas.

5.3 Outcome regressions

5.3.1 Laboratory's characteristics and management practices

Overall, the lab characteristics age, gender and number of researchers are overall not significantly associated with the different management dimensions (see Table 5.5). The only strong significant association can be seen between operational management and the number of researchers. A ten percent increase in the number of researchers is associated with an increase 1% (0.23*log(1.10)) standardized operational management score which is comprises by the two analyzed survey questions. Furthermore, the number of researcher has a slight positive association with target management, resulting in an.increase of 0.7% (0.16*log(1.10)) standard deviations associated with a ten percent increase of number of researcher (significant at a 10% level). Additionally, an increase of one year in lab age is associated with a 0.11 decrease of standardized target management score (significant at a 10% level).

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Management score			Operational management	Performance management	Target management	People management
Gender (male)			0.068	0.179	0.153	0.039	-0.093
			(0.183)	(0.162)	(0.125)	(0.126)	(0.114)
Lab age			-0.013	-0.011	-0.008	-0.011*	0.001
			(0.011)	(0.007)	(0.007)	(0.006)	(0.007)
Number of researcher (log)			0.230	0.230**	0.049	0.160*	0.117
			(0.149)	(0.093)	(0.081)	(0.087)	(0.117)
Biology		-0.254	-0.203	0.401*	0.008	-0.174	-0.277
		(0.320)	(0.333)	(0.212)	(0.203)	(0.193)	(0.245)
Genetics		0.250	0.294	0.231	0.209	0.201	0.040
		(0.255)	(0.271)	(0.214)	(0.182)	(0.160)	(0.157)
Interdisciplinary research		0.319	0.368	-0.051	0.231	0.164	0.307
		(0.341)	(0.360)	(0.320)	(0.229)	(0.249)	(0.226)
Psychiatry and Radiology		0.360	0.404	0.370	0.306*	0.213	0.127
		(0.245)	(0.256)	(0.240)	(0.167)	(0.158)	(0.171)
Constant	0.012	-2.511**	-2.809**	-3.384***	-0.769	-1.591**	-2.076***
	(0.093)	(1.030)	(1.085)	(0.752)	(0.790)	(0.766)	(0.590)
R-squared	0.000	0.096	0.125	0.201	0.077	0.163	0.133
Adjusted R-squared	0.0	0.038	0.043	0.126	-0.010	0.084	0.051
N	117	117	117	117	117	117	117
Noise Controls		X	X	X	X	X	X

* p<0.1, ** p<0.05, *** p<0.01

Table 5.5: Relationship between laboratory characteristics and management quality
Column 1 to 3 has the overall management score across all 16 management practices as dependent variable. Columns 4 to 7 have the four different management dimensions as dependent variables: Operational management is the average score across the first two questions of the survey, performance management represents the average of questions 3 to 6, target management of questions 7 to 11 and people management of questions 12 to 16. The noise controls include willingness to reveal information, duration of the interview, the analyst.

5.3.2 Laboratory's output and management practices

The main regression model for this study, with the three research output variables as dependent variables and the z-management score as main independent variable can be seen in Table 5.6. The measurement for productivity, the publication per year of the PI being last author, can be seen in column one to three. There seems to be no significant relationship with the standardized management score. Only, the number of researcher and the approximated research field (see Chapter 4) of Psychiatry and Radiology have a strong positive relationship with the number of publication. For instance, an increase of ten percent in the number of researcher is associated with an increase of 0.12 (2.88*log(1.10)) more publications per year (column 3 in Table 5.6). Also, a research lab being in the field of psychiatry and radiology seems to have 2.28 publications more per year compared to the other fields.

As can be seen in column four to six, effective management practices are associated with more citations per paper in which the principal investigator is the last author. Without controlling for research fields and other lab characteristics, I do not find a significant relationship (column 4 in Table 5.6). After controlling for research fields, I find that a one-standard-deviation increase of the z-management score (1 standardized and 0.477 unstandardized) is significantly associated with 2.03 more citations per paper (11 % relative to the unconditional mean, column 6 Table 5.6). Overall, there seems to be strong, significant disparities in the citation count across the control variables. For instance, the number of researcher has a very significant and high magnitude association with 0.32 (7.777*log(1.10)) citations more per publication per ten percent increase of number of researchers. Also lab age and being a male PI seems to be significant associated with the number of citations. Lastly, it seems that the listed approximated fields of Biology, Genetics, and Psychiatry and Radiology seem to have significantly less citations per paper than the reference field of Systems Biology. Being in the field is associated with 8.75 to 10.75 fewer citations per paper. The significant relationship between management score and citations per paper is also robust across different controls. I ran the same regression with replacing the research areas identified through the clustering,first, with the institutions (see

Appendix A.1) and second, with the departments of the research labs (see Appendix section A.2). Both regressions, show a strong and significant relationship between an increase of one standard deviation in management score being associated with an increase of 1.84 citations per paper (p<5%) for the institutional analysis and an increase of 2.31 citation per paper for the departmental analysis (p<5%) (see Appendix A)

The outcome variable of funding per researcher and year has no significant and a very small relationship with management (see column seven to nine in Table 5.6). Only the number of researcher has a negative, significant relationship with the funding per researcher as well as being in the field of interdisciplinary medical research. Being a researcher lab conducting interdisciplinary research is associated with a decrease of 33.4% (-0.334*100) in funding per researcher (see column 9 in Table 5.6). Overall, the Adjusted R-square has a high value of 0.83, meaning that 83% of the variation is explained by this regression, most of it by the number of researcher.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	Publication per year (PI last author)			Citations per paper (PI last author)			Funding per year/researcher		
Management score	0.038	-0.026	-0.017	1.087	1.846*	2.030**	0.021	0.044	0.052
	(0.278)	(0.259)	(0.265)	(1.003)	(0.968)	(0.933)	(0.055)	(0.047)	(0.045)
Number of researcher (log)	2.826***	2.879***	2.858***	7.777***	8.052***	7.777***	-1.457***	-1.440***	-1.450***
	(0.459)	(0.533)	(0.511)	(1.831)	(1.909)	(1.707)	(0.126)	(0.124)	(0.124)
Lab age			0.012			0.248*			0.009
			(0.035)			(0.138)			(0.006)
Gender (male)			0.196			4.679*			-0.137
			(0.475)			(2.374)			(0.092)
Biology		0.838	0.808		-10.238**	-10.747**		0.237	0.185
		(1.015)	(1.058)		(4.826)	(4.577)		(0.170)	(0.166)
Genetics		0.336	0.243		-7.626*	-9.573**		-0.067	-0.123
		(0.695)	(0.776)		(3.853)	(3.777)		(0.131)	(0.144)
Interdisciplinary research		-0.591	-0.611		-7.909	-8.291		-0.292**	-0.334**
		(0.825)	(0.847)		(5.089)	(5.226)		(0.134)	(0.145)
Psychiatry and Radiology		2.320***	2.283**		-8.181*	-8.753**		0.071	0.026
		(0.872)	(0.937)		(4.381)	(4.103)		(0.154)	(0.150)
Published paper in last 5 year				0.013	0.042	0.029			
				(0.105)	(0.097)	(0.087)			
Constant	-1.887**	-1.603	-1.810	1.087	26.771**	22.186*	5.002***	5.725***	5.606***
	(0.905)	(2.563)	(2.781)	(3.929)	(11.232)	(12.567)	(0.278)	(0.583)	(0.568)
R-squared	0.327	0.408	0.410	0.163	0.302	0.352	0.823	0.843	0.850
Adjusted R-squared	0.316	0.358	0.348	0.141	0.237	0.277	0.819	0.829	0.833
N	117	117	117	117	117	117	108	108	108
Noise Controls		X	X		X	X		X	X

* p<0.1, ** p<0.05, *** p<0.01

Table 5.6: Relationship between laboratory outcomes and management quality

The three dependent variables serve as a proxy for lab's success. Management score is the standardized average across all 16 management practice scores from the survey data. Lab age is the time since the foundation of the research lab and the independent variables biology to dermatology and radiology should control for the research field. The noise controls include willingness to reveal information, duration of the interview and the analyst.

5.3.3 Splitting young and old labs - citations and management practices

Splitting the data set at the median lab age (15 years), results in two groups: young (below 15 years), and old (equal or older than 15 years) labs. Using the same control variables for the regressions as in column 6, Table 5.6. Compared to the baseline of labs of all ages, the effect size of the standardized management score increases for the younger group by 50% to 3.009 ($p< 0.05$). An increase of one standard deviation in management score is associated with three citations per paper more, an increase of 0.17% relative to the unconditional mean of 17.63 citation per paper for young labs. This significant relationship seems to disappear for the older labs, which effect size decreases significantly to 0.852 and represents no significant effect anymore.

	(1)	(2)	(3)
	Citations per Paper		
	Baseline	Young labs	Old labs
Management score	2.030**	3.009**	0.852
	(0.933)	(1.308)	(1.610)
Constant	22.186	15.005	21.485
	(12.567)	(17.347)	(17.686)
R-squared	0.352	0.377	0.402
Adjusted R-squared	0.277	0.221	0.235
N	117	61	56
Noise control	X	X	X
Controls	X	X	X

* p<0.1, ** p<0.05, *** p<0.01

Table 5.7: Relationship between citations and management quality relative to lab's age
Column 1 is the same regression than in Table 5.6 and column 2 and 3 is the same regression split into two subgroups: young research labs (<15 years) and old labs (≥ 15 years). The control variables are the same than in Table 5.6, controlling for number of researcher, gender, lab age, lab areas and published papers in the last five years. The noise controls include willingness to reveal information, duration of the interview, the analyst.

5.3.4 Differences in laboratory's collaboration structure

Another dimension of this research is to look at different laboratory structures. Specifically, the labs were categorized according to their collaboration policy on projects. As can be seen in Figure 5-2, almost two thirds (63%) of the interviewed laboratories employ collaborative projects. Management is slightly higher for research labs with collaboration with a mean of 3.21 versus 3.03. Labs with collaborative projects have on average 18.5% more publications per year comparing 4.47 with 3.77 publication per year in both groups. The citations per paper are, on average, higher in research labs with collaborative projects (19.83 compared to 15.69 citations per paper). While all three measures have a higher magnitude in labs having implemented collaborative project structures, the error bars for all measurements are very significant. None of the measurements are statistically significant as the confidence intervals overlap.

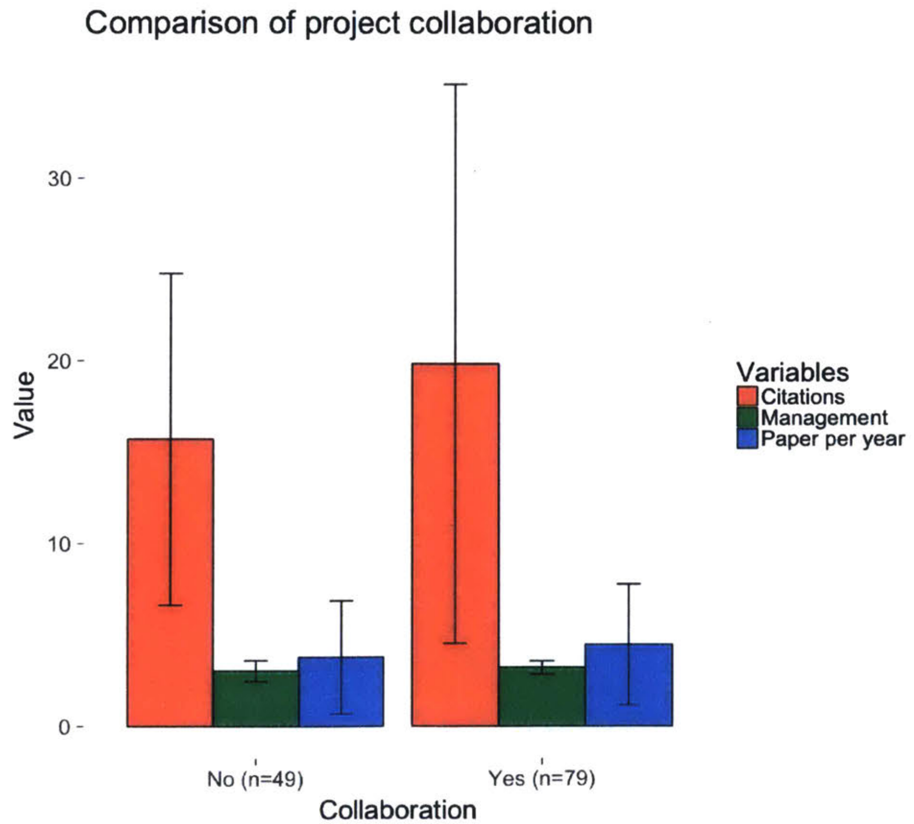

Figure 5-2: Collaboration effort

Research labs are categorized in having collaborative projects, meaning that at least two researchers are responsible for a project, or non-collaborative projects, meaning that only one researcher is responsible for it's success. The two groups are compared across the overall management score and the two outcome variables, citations per publication per researcher and number of publications per year per researcher. Depicted are the averages with associated error bars.

THIS PAGE INTENTIONALLY LEFT BLANK

Chapter 6

Discussion of Results

6.1 Understanding the management score

The Section 5.2.2 shows interesting variations of the management score across domains, institutions and areas. First, performance management is scored significantly higher on average than target management, which could indicate that in medical research there are well established management processes in the sampled labs to track the performance of ongoing research projects. However, there might be a lack of leveraging management practices to set new targets and as the nature of research goes, defining targets might be harder in the first place. Moreover, operational management has a smaller distribution than people management, which could mean that the management practices for people management are less defined or the execution has more variance. A potential reason could be that every PI has a very different opinion on how to manage and mentor their researchers and there is more subjectivity involved than in operational management.

The higher variation within institutions than between them, could mean that overall the PI's individual management capabilities are a far more important factor than institutional affiliation. The variation between the institutions however, could potentially reflect that some institutions, such as the Dana-Farber Cancer Institute, started to introduce managerial guidelines to their PIs which other institutions do not have. It could further reflect the fact that each institution houses different research domains as assumed in the previous chapter. These results could also reflect that the initial efforts of institutions to teach their staff management practices are at an early stage and until now, and do not outweight the importance of personal management practices initiated by the PI independently of the institution.

Even a stronger disparity of between and within variation is found in the different research areas. Even though the fields seem to be quite different (i.e. Radiology and Psychiatry versus Biology), these results suggest that the fields have no big impact on differences in management score. The strong within research field variation could support the results from the institutional variation in that the individual capabilities of the PI are more important than the field and thus the management score can vary strongly even though the PIs are in comparable settings when it comes to resources and managerial requirements.

Interpreting the result in Section 5.3.1 offers a more detailed perspective into the management score. The relationship between the laboratory characteristics and the management practices reveal that almost none of the main characteristics (i.e. gender, lab age and number of researchers) seem to have a significant relationship with the different management dimensions. However, an interesting and very significant relationship is between the number of researchers and operational management. This makes intuitive sense as operational management questions are about how to allocate and manage various resources across the staff and experiments. Increasing the number of people in the lab and with that, the number of projects, can lead to a higher demand for managing these resources in a standardized and transparent way - one of the criteria to obtain a high score in the operational management dimension. Interestingly, there is no significant difference when it comes to gender. Lab age does show a small negative relationship with target management while the number of researchers has a positive association with it. It could be that the more mature a lab becomes, to some extent it also becomes complacent or has the luxury of not having to set very specific goals anymore to sustain itself. For instance, a young lab seems to have a harder time to secure enough funding which in return depends on publications. Thus, young labs have very clear goals when to publish what research. Additionally, having more people in the research lab might require better target management, or better target management enables to have more people. The first version makes intuitively more sense: with an increasing number of researchers, the number of projects and associated milestones/targets can grow exponentially. In order to keep track, the PI might implement better target management practices such as managing the interconnection of targets and the time horizons of targets (question 10 and 11 in the survey) better. Also interesting is the lack of significant association between lab age and management score. I would have expected that older labs have a higher management score as they have much experience and survived the tenure process. The absence of this significant relationship can have multiple reasons. Perhaps, there is simply no learning effect over time. One reason could be that there is no feedback loop. Most PIs simply manage their labs as they have learned it from their previous supervisor. Since the hierarchies are strong within the lab and there is little knowledge of what best practices is to run a lab, there is very little feedback the PI can get from within or outside the lab. It is one of the reasons this study is so important. There could also be two conflicting effects which balance each other out. For one, some PIs might get better managing their laboratories over time while others, even without great management skills have been able to lead a successful lab, become complacent over time and their management of projects and people decreases again over time. Both effects could negate other impacts (splitting the labs into three equal grous of young, middle and old shows the highest average management scores for the middle group). Either way, future research should try to shed light on the drivers of good management.

6.2 The main regression results

Research laboratories with higher management scores have more citations per paper on average. On the other hand, management had no significant association with the number of publications or the level of funding per lab. A potential reason for these results could be that better management in terms of operations, performance, target and people management, might not be associated with a higher quantity of publications or shorter cycles between publications but rather with more rigorous research, which is then cited more often. Potentially, while the time cycles of publications stay in-tact, the better managed research labs can use the time to make the analysis more rigorous such as run an additional experiement to substantiate the results. This can lead to a publication in a more highly recognized journal and/or being perceived as higher quality research by colleagues to build up their own research on or refer to. Funding has a different dynamic, it is a necessity, yet once the need to finance the lab's members and equipment demands is satisfied, accumulating more funding does not necessarily help in conducting better research. Further, the structure of research grants often does not allow to re-apply for certain funding offers and funding is strictly tied to the actual resources needed by the research project.

The relationship between management and citations per paper is more significant for younger than for older labs. As discussed in Section 2.1.1, the citation count is an imperfect, yet the best possible proxy for research quality. Other researcher cite publications for multiple other reasons than quality such as if there is current public interest in the topic, if the publication journal is highly ranked (independently of the quality of the actual research) and last but not least, how well known the research lab and PI is. Especially the last point might be the reason why more mature labs have not a strong relationship with management in regards to citation counts as the reputation of the lab and PI becomes a dominating factor. Another point is the selection bias, as only the best principal investigators in the best labs survive the long tenure track process and stay successful in academia over time. The 56 labs in my sample are therefore pre-selected which make it harder to disentangle the management impact from this selection effect. However, younger laboratories and PIs should have a smaller selection effect as they are less known in the field and e.g. their fame does not impact the citation count of their publications to a significant extend yet. Thus, instead of their name and reputation, younger labs do rely more on the actual quality of work which potentially is positively impacted by having better management practices. Since it takes time to gain a reputation, younger labs might can or have to focus on these management practices to stay competitive. Furthermore, they do not have as much experience as their older counterparts, meaning that many e.g. research techniques might be new to them. Certain management practices could potentially be especially beneficial for younger labs, such as operational management, than older labs. Since younger labs cannot rely on experience they must employ superior management practices. Overall, these effects could lead to the described results showing that the effect of management seems to be quite important for young labs.

6.3 Effects of collaboration

The comparison of project collaboration is intuitively appealing but does not yet lead to any conclusive results. All variables in terms of management, citations per paper and paper per researcher are higher for the collaborative group than the labs without collaboration. While this effect is quite significant, the variation in the observed sample is very high and given the small sample size of splitting 117 observation, does not yield any significant results. It is interesting that not only the productivity is higher with more publications per year, but also the citation per paper. One reason could be that having multiple people on one projects helps to overcome bottlenecks, an isolated researcher might face when e.g. waiting on the help of a second person who is less invested in the project. The overall time could be more efficiently used and yield to more publications per year (controlling for the number of researcher in a lab seems to result in having almost the same number of publication per researcher in both settings). Also, there can be many reasons why the citations as a proxy of research quality is higher for collaborative projects. For instance, having multiple researchers on a project serves as a peer review system only publishing well-designed research studies, helps to better formulate the publication or is better to employ sophisticated research methods as the skill and knowledge pool for the project is larger. Previous studies have already found a positive correlation between collaboration and higher citations counts and came to the conclusion that it is most likely due to better research quality Figg et al. (2006). Especially in recent years, medical research became more complex with research projects demanding a broad range of expertise ranging from biology to chemistry and data science. Further research has to be conducted in order to understand this trend and its implications for the organizational structures of the lab better.

6.4 Limitations

First, this is an exploratory study applying a proven research tool, the World Management Survey, in a new context, biomedical research. Until now, the survey has not been cognitively tested yet and relies on other validation measures (such as cross-validation of the interviewer and listener). In a next step, cognitive testing is planned.

Second, this is not a causal study and it might suffer from reversed causality and endogeneity issues. For instance, I cannot observe variables such as the PI's ability or cognitive skills. This can lead to a correlated omitted variable bias. For example, the intelligence of the PI might not only affect management practices, but can also directly lead to more and highly-cited publications. Additionally, reversed causality cannot be ruled out. Better publications can lead to more funding, which again can lead to the PI affording a lab manager or more researcher which causes improvements in the lab management. In a subsequent study, I plan to conduct a field experiment to establish causal relations (e.g., in the spirit of Bloom et al. (2013)).

Third, my proxies for research quality are imperfect. I want to understand what the impact of management on research quality is. While I use the number of publications, citations and funding as legitimate proxies for the "success" of the research lab, these are not necessarily perfect proxies for true research quality. In a subsequent step, I might want to look at other outcome variables of research (e.g. measures for novelty) which better depict what makes research innovative and 'good'.

Fourth, I only interviewed PIs, as I felt that they are the best interview partners as they have the most decision power in the research laboratory. To better control for the validity of their statements and gather data from a different perspective, it would be helpful to interview other researchers and staff within the same laboratory as well.

Fifth, while I have a high response rate of 33% and my results and interviews support the conjecture that only few laboratories actively think about management practices, one could be concerned that the interviewed PIs might not be representative of the full population. However, I believe that, if anything, labs in my sample are more likely to care about management as they were willing to participate in this survey about management practices. Thus, my results are more likely to present an upper bound when it comes to the average quality management practice in medical labs.

THIS PAGE INTENTIONALLY LEFT BLANK

Chapter 7

Conclusion

7.1 Results of this study

This study is one of the first in-depth analyses of management practices in medical research. It is the first of its kind to employ the World Management Survey in the context of academic research and collect meaningful data from a relatively large sample of research labs. The results of this study show that this new survey tool allows to collect a range of potentially important data items, which is valuable for academic research shedding light into the unknowns of management in medical research.

The results of the study are consistent with management playing an important role in medical research. Better management practices are associated with more citations per publication, a widely-recognized proxy for research quality.

Furthermore, the study finds that this effect is stronger and more significant in younger compared to older laboratories. Additionally, it seems that neither the size of the research lab, the lab age, nor the gender of the PI determines management quality. Rather, without being able to control for e.g. the intelligence of the PI, the comparisons of management quality across institutions and research areas suggest that management quality within a research lab seems to be driven by the individual capabilities of the PI, instead of his affiliation or research field. While this study finds first signs of the importance of collaborative structures in research laboratories, none of the results are conclusively significant. However, it does seem that on average collaborative research labs have a higher citation count, publish more, and have a slightly higher management quality.

7.2 Implications

This study is an important first step towards a better understanding of the drivers of "good" or "successful" research. Future research should build upon these efforts by leveraging the validated tool of the World Management Survey, conducting interviews with a larger sample and investigating further interesting aspects of this study such as the effect of collaboration and management on the quality of research. These consecutive studies could be used to identify effective management practices and organizational structures which could be tested in a field experiment to obtain causal relationships. The overarching goal of this research initiative should be to inform leaders and members of medical research laboratories (and in the long-run academic labs in general) on how their lab can be managed and structured in a more effective way. Furthermore, it should inform policy makers, universities, funding institutions such as the NIH and NSF as well as scientific journals if and what systematic changes the current funding, recognition and educational structures biomedical research needs to correctly incentivize and align future research efforts.

In the long run, this research could not only help to make more effective use of large public investments through research organizations but furthermore, can have a direct impact on the future of health care. Tracing back a question from the introduction: Would better management practices produce better science? My current results show that there is a significant link between both and it is more than worth the effort to invest further research into this promising research endeavour.

Bibliography

R. America. U.s. investments in medical and health research and development. *TEConomy Partners, LLC*, 2017.

R. C. Barnett, P. Carr, A. D. Boisnier, A. Ash, R. H. Friedman, M. A. Moskowitz, and L. Szalacha. Relationships of gender and career motivation to medical faculty members' production of academic publications. *Academic Medicine*, 1998.

J. B. Barney. Resource-based theories of competitive advantage: A ten-year retrospective on the resource-based view. *Journal of management*, 27(6):643–650, 2001.

Y. Baruch and D. T. Hall. The academic career: a model for future careers in other sectors? *Journal of Vocational Behavior*, 64(2):241–262, 2004.

M. C. Becker. Organizational routines: a review of the literature. *Industrial and corporate change*, 13(4):643–678, 2004.

C. G. Begley and J. P. Ioannidis. Reproducibility in science: improving the standard for basic and preclinical research. *Circulation research*, 116(1):116–126, 2015.

N. Bloom. *Measuring and Explaining Management Practices Across Firms and Countries*, volume 122. 2007.

N. Bloom and J. Van Reenen. Measuring and explaining management practices across firms and countries. *The Quarterly Journal of Economics*, 122(4):1351–1408, 2007.

N. Bloom, S. Dorgan, J. Dowdy, and J. Van Reenen. Management practice and productivity: Why they matter. *Management Matters*, 10, 2007.

N. Bloom, C. Genakos, R. Martin, and R. Sadun. Modern management: good for the environment or just hot air? *The Economic Journal*, 120(544):551–572, 2010.

N. Bloom, C. Genakos, R. Sadun, and J. Van Reenen. Management practices across firms and countries. *The Academy of Management Perspectives*, 26(1):12–33, 2012a.

N. Bloom, R. Sadun, and J. Van Reenen. Does management really work? how three essential practices can address even the most complex global problems. *Harvard Business Review: HBR*, 90(11):76–82, 2012b.

N. Bloom, R. Sadun, and J. Van Reenen. The organization of firms across countries. *The quarterly journal of economics*, 127(4):1663–1705, 2012c.

N. Bloom, B. Eifert, A. Mahajan, D. McKenzie, and J. Roberts. Does management matter? evidence from india *. *The Quarterly Journal of Economics*, 128(1):1–51, 2013. doi: 10.1093/qje/qjs044. URL `http://dx.doi.org/10.1093/qje/qjs044`.

N. Bloom, R. Lemos, R. Sadun, and J. Van Reenen. Does management matter in schools? *The Economic Journal*, 125(584):647–674, 2015.

L. Bornmann and H.-D. Daniel. What do citation counts measure? a review of studies on citing behavior. *Journal of documentation*, 64(1):45–80, 2008.

H. K. Bowen and F. Gino. The whitesides lab. *Harvard Business School Case*, 606-064, 2006.

H. K. Bowen, A. Kazaks, A. Muir-Harmony, and B. LaPierre. Langer lab, the: Commercializing science. 2004.

J. Brainard. Nih casts critical eye on how it gives grants. *The Chronicle of Higher Education*, 2007.

N. Carayol and M. Matt. Does research organization influence academic production?: Laboratory level evidence from a large european university. *Research Policy*, 33(8):1081–1102, 2004.

Z. Chinchilla-Rodríguez, M. Benavent-Pérez, F. Moya-Anegón, and S. Miguel. International collaboration in medical research in latin america and the caribbean (2003–2007). *Journal of the Association for Information Science and Technology*, 63(11):2223–2238, 2012.

F. S. Collins and L. A. Tabak. Nih plans to enhance reproducibility. *Nature*, 505(7485):612, 2014.

L. J. Cronbach. Coefficient alpha and the internal structure of tests. *psychometrika*, 16(3): 297–334, 1951.

C. K. De Dreu and M. A. West. Minority dissent and team innovation: The importance of participation in decision making. *Journal of applied Psychology*, 86(6):1191, 2001.

A. M. Diamond. The money value of citations to single-authored and multiple-authored articles. *Scientometrics*, 8(5-6):315–320, 1985.

J. A. DiMasi, H. G. Grabowski, and R. W. Hansen. Innovation in the pharmaceutical industry: new estimates of r&d costs. *Journal of health economics*, 47:20–33, 2016.

P. Dong, M. Loh, and A. Mondry. The" impact factor" revisited. *Biomedical digital libraries*, 2 (1):7, 2005.

L. Egghe and R. Rousseau. *Introduction to informetrics: Quantitative methods in library, documentation and information science*. Elsevier Science Publishers, 1990.

K. M. Eisenhardt and J. A. Martin. Dynamic capabilities: what are they? *Strategic management journal*, pages 1105–1121, 2000.

Elsevier. *Scopus - bibliographic database*. 2017. doi: https://www.scopus.com/search/.

E. Ernø-Kjølhede. Managing researchers. *Science and Public Policy*, 28(1):49–55, 2001.

W. D. Figg, L. Dunn, D. J. Liewehr, S. M. Steinberg, P. W. Thurman, J. C. Barrett, and J. Birkinshaw. Scientific collaboration results in higher citation rates of published articles. *Pharmacotherapy: The Journal of Human Pharmacology and Drug Therapy*, 26(6):759–767, 2006.

B. S. Frey and S. Neckermann. Awards: A view from economics. *The Economics of Ethics*, pages 73–88, 2009.

J. Friedman and J. Silberman. University technology transfer: do incentives, management, and location matter? *The Journal of Technology Transfer*, 28(1):17–30, 2003.

E. Garfield et al. The impact factor. *Current contents*, 25(20):3–7, 1994.

A. Gazni and M. Thelwall. The long-term influence of collaboration on citation patterns. *Research Evaluation*, 23(3):261–271, 2014.

N. Ghaffarzadegan, J. Hawley, R. Larson, and Y. Xue. A note on phd population growth in biomedical sciences. *Systems research and behavioral science*, 32(3):402–405, 2015.

D. K. Ginther, W. T. Schaffer, J. Schnell, B. Masimore, F. Liu, L. L. Haak, and R. Kington. Race, ethnicity, and nih research awards. *Science*, 333(6045):1015–1019, 2011.

J. Golden and F. V. Carstensen. Academic research productivity, department size and organization: Further results, comment. *Economics of Education Review*, 11(2):153–160, 1992.

L. O. Gostin, L. A. Levit, S. J. Nass, et al. *Beyond the HIPAA privacy rule: enhancing privacy, improving health through research*. National Academies Press, 2009.

D. K. Harman and E. M. Voorhees. Trec: An overview. *Annual review of information science and technology*, 40(1):113–155, 2006.

Z.-L. He, X.-S. Geng, and C. Campbell-Hunt. Research collaboration and research output: A longitudinal study of 65 biomedical scientists in a new zealand university. *Research Policy*, 38(2):306–317, 2009.

F. Hecht, B. K. Hecht, and A. A. Sandberg. The journal "impact factor": a misnamed, misleading, misused measure. *Cancer genetics and cytogenetics*, 104(2):77–81, 1998.

F. Hillen. Determinants of successful medical research: An empirical study of management practices in research laboratories. *Bachelorthesis*, 2016.

J. E. Hirsch. An index to quantify an individual's scientific research output. *Proceedings of the National academy of Sciences of the United States of America*, 102(46):16569, 2005.

C. Ichniowski, K. Shaw, and G. Prennushi. The effects of human resource management practices on productivity: A study of steel finishing lines. *The American Economic Review*, pages 291–313, 1997.

D. Jindal-Snape and J. B. Snape. Motivation of scientists in a government research institute: Scientists' perceptions and the role of management. *Management Decision*, 44(10):1325–1343, 2006.

R. Kassab and J.-C. Lamirel. Feature-based cluster validation for high-dimensional data. In *Proceedings of the 26th IASTED International Conference on Artificial Intelligence and Applications*, pages 232–239. ACTA Press, 2008.

P. Kline. *Handbook of psychological testing*. Routledge, 2013.

A. V. Kulkarni, J. W. Busse, and I. Shams. Characteristics associated with citation rate of the medical literature. *PloS one*, 2(5):e403, 2007.

M. L. Lengnick-Hall, C. A. Lengnick-Hall, L. S. Andrade, and B. Drake. Strategic human resource management: The evolution of the field. *Human resource management review*, 19(2): 64–85, 2009.

D. P. Lepak, H. Liao, Y. Chung, and E. E. Harden. A conceptual review of human resource management systems in strategic human resource management research. In *Research in personnel and human resources management*, pages 217–271. Emerald Group Publishing Limited, 2006.

Z. Li and Y.-S. Ho. Use of citation per publication as an indicator to evaluate contingent valuation research. *Scientometrics*, 75(1):97–110, 2008.

H. J. Lowe and G. O. Barnett. Understanding and using the medical subject headings (mesh) vocabulary to perform literature searches. *Jama*, 271(14):1103–1108, 1994.

C. Macilwain. What science is really worth: spending on science is one of the best ways to generate jobs and economic growth, say research advocates. but as colin macilwain reports, the evidence behind such claims is patchy. *Nature*, 465(7299):682–685, 2010.

H. Moses, E. Dorsey, D. Matheson, and S. Thier. Financial anatomy of biomedical research. *JAMA*, 294(11):1333–1342, 2005. doi: 10.1001/jama.294.11.1333. URL +http://dx.doi.org/10.1001/jama.294.11.1333.

C. Nemeth and B. Nemeth-Brown. Better than individual? the potential benefits of dissent and diversity. *P. B, Paulus, and BA Nijstad,(Ed.), Group creativity*, pages 63–84.

P. Nieminen, J. Carpenter, G. Rucker, and M. Schumacher. The relationship between quality of research and citation frequency. *BMC Medical Research Methodology*, 6(1):42, 2006.

NIHReporter. *National Institute of Health*. 2017. doi: https://report.nih.gov/index.aspx/.

D. Nobelius. Towards the sixth generation of r&d management. *International Journal of Project Management*, 22(5):369–375, 2004.

NSF. *About the National Science Foundation*, 2017. https://www.nsf.gov/about/ [Accessed: 04/19/2017].

S. Numprasertchai and B. Igel. Managing knowledge through collaboration: multiple case studies of managing research in university laboratories in thailand. *Technovation*, 25(10): 1173–1182, 2005.

F. Radicchi, S. Fortunato, and C. Castellano. Universality of citation distributions: Toward an objective measure of scientific impact. *Proceedings of the National Academy of Sciences*, 105 (45):17268–17272, 2008.

J. Rigby and J. Edler. Peering inside research networks: Some observations on the effect of the intensity of collaboration on the variability of research quality. *Research policy*, 34(6): 784–794, 2005.

A. M. Sapienza. *Managing scientists: leadership strategies in scientific research (Vol. 2)*. John Wiley & Sons, INC., Publication., Boston, Massachusetts, 2004.

C. C. Sarli, E. K. Dubinsky, and K. L. Holmes. Beyond citation analysis: a model for assessment of research impact. *Journal of the Medical Library Association: JMLA*, 98(1):17, 2010.

P. E. Stephan. *How economics shapes science*, volume 1. Harvard University Press Cambridge, MA, 2012.

P. F. Svider, K. M. Mauro, S. Sanghvi, M. Setzen, S. Baredes, and J. A. Eloy. Is nih funding predictive of greater research productivity and impact among academic otolaryngologists? *The Laryngoscope*, 123(1):118–122, 2013.

R. Tibshirani, G. Walther, and T. Hastie. Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(2):411–423, 2001.

N. J. Van Eck, L. Waltman, A. F. van Raan, R. J. Klautz, and W. C. Peul. Citation analysis may severely underestimate the impact of clinical research as compared to basic research. *PloS one*, 8(4):e62395, 2013.

E. West. Management matters: the link between hospital organisation and quality of patient care. *BMJ Quality & Safety*, 10(1):40–48, 2001.

C. Woolston. Graduate survey: a love-hurt relationship. *Nature*, 550(7677):549–552, 2017.

S. Wuchty, B. F. Jones, and B. Uzzi. The increasing dominance of teams in production of knowledge. *Science*, 316(5827):1036–1039, 2007.

THIS PAGE INTENTIONALLY LEFT BLANK

Appendix A

Tables

A.1 Main regression - controlling with research institution

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	Publication per year (PI last author)			Citations per paper (PI last author)			Funding per year/researcher		
Management score	0.038	0.038	0.036	1.087	1.690*	1.837**	0.021	0.032	0.038
	(0.278)	(0.282)	(0.284)	(1.003)	(0.876)	(0.877)	(0.055)	(0.049)	(0.043)
Number of researcher (log)	2.826***	2.850***	2.875***	7.777***	8.233***	8.305***	-1.457***	-1.442***	-1.470***
	(0.459)	(0.475)	(0.507)	(1.831)	(1.986)	(1.914)	(0.126)	(0.128)	(0.123)
Lab age			0.021			0.190			0.008
			(0.033)			(0.139)			(0.005)
Gender (male)			-0.026			5.370**			-0.142
			(0.545)			(2.332)			(0.095)
Brigham Hospital			-0.388			-8.298***			0.142
			(0.890)			(2.814)			(0.190)
Dana - Farber Center			-0.040			1.680			-0.261
			(0.907)			(5.580)			(0.172)
Harvard University			-0.664			-5.101*			-0.254*
			(0.865)			(2.909)			(0.147)
MGH			0.386			-5.358*			-0.255*
			(0.936)			(3.058)			(0.140)
Others			-0.330			-2.696			0.006
			(0.765)			(2.951)			(0.265)
Published paper in last 5 year				0.013	0.008	-0.003			
				(0.105)	(0.100)	(0.096)			
Constant	-1.887**	-1.623	-1.892	1.087	23.056**	18.776	5.002***	5.570***	5.408***
	(0.905)	(2.774)	(3.078)	(3.929)	(11.124)	(11.571)	(0.278)	(0.600)	(0.566)
R-squared	0.327	0.328	0.348	0.163	0.242	0.339	0.823	0.829	0.857
Adjusted R-squared	.3155577	.2979636	.2708024	.1410512	.2006989	.2544134	.8193441	.8208295	.8385336
N	117	117	115	117	117	115	108	108	107
Noise Controls		X	X		X	X		X	X

\footnotesize * p<0.1, ** p<0.05, *** p<0.01

Table A.1: Relationship: laboratory outcomes and management with institutions as controls
Column 1 to 3 has the overall management score across all 16 management practices as dependent variable. Columns 4 to 7 have the four different management dimensions as dependent variables: Operational management is the average score across the first two questions of the survey, performance management represents the average of questions 3 to 6, target management of questions 7 to 11 and people management of questions 12 to 16. The noise controls include willingness to reveal information, duration of the interview, the analyst.

A.2 Main regression - controlling with research departments

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	paper_year	paper_year	paper_year	Citations per paper (PI last author)	Citations per paper (PI last author)	Citations per paper (PI last author)	llmoney	llmoney	llmoney
Management score	0.038	-0.036	-0.025	1.087	2.200**	2.309**	0.021	0.050	0.050
	(0.278)	(0.298)	(0.297)	(1.003)	(1.098)	(1.059)	(0.055)	(0.061)	(0.060)
Number of researcher (log)	2.826***	2.955***	2.908***	7.777***	8.340***	8.360***	-1.457***	-1.439***	-1.449***
	(0.459)	(0.478)	(0.460)	(1.831)	(2.198)	(2.032)	(0.126)	(0.119)	(0.119)
Lab age			0.021			0.038			0.005
			(0.033)			(0.147)			(0.006)
Gender (male)			0.279			5.328*			-0.090
			(0.681)			(2.822)			(0.108)
Cell Biology		0.150	0.228		-2.936	-4.040		-0.084	-0.010
		(1.563)	(1.655)		(7.847)	(7.190)		(0.262)	(0.280)
Dermatology		3.113	3.214		-11.463	-12.106		-0.336**	-0.275
		(2.007)	(2.079)		(8.446)	(7.626)		(0.165)	(0.185)
Developmental Biology		-2.246	-2.105		-5.479	-4.534		-0.311**	-0.327**
		(1.386)	(1.528)		(7.334)	(5.706)		(0.130)	(0.159)
Genetics and Complex Diseases		0.028	0.019		1.700	0.555		0.036	0.072
		(1.148)	(1.308)		(7.469)	(7.097)		(0.260)	(0.272)
Immunology and Infectious Diseases		-1.217	-0.978		-12.745	-12.355*		-0.491***	-0.430**
		(1.146)	(1.229)		(9.493)	(7.254)		(0.153)	(0.186)
Medicine		-0.843	-0.746		-3.914	-4.736		-0.175	-0.116
		(1.075)	(1.229)		(5.990)	(4.847)		(0.150)	(0.174)
Microbiology and Immunobiology		0.097	0.156		-9.245	-10.973*		-0.358*	-0.281
		(1.953)	(2.027)		(6.962)	(6.561)		(0.201)	(0.234)
Neurobiology		-0.298	-0.217		-6.550	-6.632		-0.039	-0.020
		(2.272)	(2.317)		(6.904)	(6.468)		(0.266)	(0.262)
Neurology		-1.920	-1.887		-5.733	-8.173		-0.438**	-0.349
		(1.226)	(1.418)		(6.457)	(5.969)		(0.185)	(0.224)
Obstetrics Gynecology and Repro. Bio.		-1.644	-1.391		-2.428	0.650		-0.155	-0.172
		(1.400)	(1.430)		(8.273)	(8.378)		(0.261)	(0.282)
Pathology		-0.516	-0.364		-4.918	-5.178		-0.273*	-0.216
		(1.311)	(1.398)		(5.832)	(4.847)		(0.145)	(0.175)
Pediatrics		-0.034	0.135		-4.930	-4.575		-0.102	-0.058
		(1.084)	(1.184)		(5.654)	(4.682)		(0.190)	(0.214)
Psychiatry		0.766	0.856		-3.488	-4.091		0.205	0.257
		(1.523)	(1.576)		(7.344)	(6.524)		(0.256)	(0.267)
Radiology		-0.710	-0.560		-11.714	-11.997**		-0.208	-0.152
		(1.546)	(1.663)		(7.081)	(6.028)		(0.221)	(0.237)
Stem Cell and Regenerative Biology		-1.302	-1.242		-7.949	-8.911		-0.346*	-0.297
		(1.813)	(1.913)		(6.809)	(6.770)		(0.182)	(0.202)
Surgery		2.006	2.226		-5.707	-3.438		0.731	0.728
		(1.397)	(1.529)		(5.400)	(4.664)		(0.477)	(0.519)
Systems Biology		-2.642**	-2.833**		3.666	0.885		-0.299	-0.270
		(1.117)	(1.354)		(8.160)	(7.795)		(0.229)	(0.254)
Published paper in last 5 year				0.013	0.033	0.021			
				(0.105)	(0.107)	(0.101)			
Constant	-1.887**	-0.121	-0.608	1.087	24.406	22.784	5.002***	5.877***	5.758***
	(0.905)	(3.498)	(3.772)	(3.929)	(16.981)	(17.092)	(0.278)	(0.672)	(0.678)
R-squared	0.327	0.435	0.439	0.163	0.314	0.342	0.823	0.865	0.867
Adjusted R-squared	.3155577	.3025879	.2929942	.1410512	.1444748	.1613412	.8193441	.8296219	.8290126
N	117	117	117	117	117	117	108	108	108
Noise Controls		X	X		X	X		X	X

* p<0.1, ** p<0.05, *** p<0.01

Table A.2: Relationship: laboratory outcomes and management with departments as controls
Column 1 to 3 has the overall management score across all 16 management practices as dependent variable. Columns 4 to 7 have the four different management dimensions as dependent variables: Operational management is the average score across the first two questions of the survey, performance management represents the average of questions 3 to 6, target management of questions 7 to 11 and people management of questions 12 to 16. The noise controls include willingness to reveal information, duration of the interview, the analyst.

Appendix B

Medical Research Survey

Medical Research Survey

Interview Details

Research laboratory ID: ____________________________________

Name of Institution: ____

Interviewer Name: ________________________

Date (DD/MM/YY): ________________________

Time (24 hour clock): _____________________
Running interview □ Listening to interview □

Laboratory and Principal Investigator (PI) Information

a) Position: _________________

b) Tenure in post: ______

c) Tenure in institution: ______

d) Number of Researchers(PD/PhD/Student): ______

e) Number of Assistants (Research/Administrative/Technicians): ______

f) Number of lab managers: ______

g) Number of active projects in the lab: ______

h) Number of competitors: ______

Management Questions

1) Standardizations and Protocols

Test for standardized main processes in the laboratory (lab operations and experiments)

Score:

1☐ 2☐ 3☐ 4☐ 5☐

a) What tasks in your lab need to be followed up on, on a regular basis? (Ask for either operational maintenance tasks (lab manager or technician) and routine in experiment work (researcher))
b) How standardized would you say are these processes/tasks?
c) What tools and resources does the lab´s staff use (e.g. checklists) to ensure that they conduct the appropriate procedure?
d) How clear are lab´s staff members about how specific procedures should be carried out?

Score 1: Little standardization and few protocols exists (e.g. different lab´s staff have different approaches to the same action)	Score 3: Protocols have been created, but are not commonly used because they are too complicated or not communicated effectively	Score 5: Protocols are known and used by all lab´s staff and regularly followed up on through some form of monitoring or oversight (e.g. of the PI)
Low	Degree of standardization and protocols	High

2) Rationale for Standardized processes?

Tests the motivation and impetus behind changes to operations and what change story was communicated

Score:1☐ 2☐ 3☐ 4☐ 5☐

a) What is the rationale for making such standardized tasks/process improvements?
b) How often do you challenge these tasks (or processes) which need be done regularly?

Score 1: Changes were imposed out of a necessity or because other laboratories were making (similar) changes;(e.g. industry standards) rationale was not communicated or understood	Score 3: Changes were made because of financial pressure (to reduce costs)	Score 5: Changes were made to **improve overall performance**, both research-output wise and financial, with buy-in from all affected staff groups; the changes were communicated in a coherent 'change story' (to increase efficiency)
Low	Degree of "motivation" for standardization	High

Medical Research Survey

3) Continuous improvements

Tests processes for and attitudes to continuous improvement and whether learnings are captured/ documented

Score:

1□ 2□ 3□ 4□ 5□

a) How do problems get exposed and fixed?

b) How can your staff suggest improvements in your lab? Can you think of a recent example? (Probing: Do researchers or technicians come to you to suggest an improvement? How often does this happen?)

c) Talk me through the process for a recent problem?

Score 1: : No process improvements are made when problems occur	Score 3: Improvements are made in (ir)regular meetings involving all staff groups, to improve performance in their project	Score 5: Exposing problems in a structured way is integral to an individual's responsibilities and resolution involves all research teams. Exposing and resolving problems is part of a regular business process rather than being the result of extraordinary efforts
Low	Degree of continues improvement efforts	High

4) Good use of human resources

Tests processes for and attitudes to collaboration and knowledge exchange between researchers

Score:

1□ 2□ 3□ 4□ 5□

a) With respect to your staff, what happens when one researcher needs the expertise of another one?

b) What kind of procedures do you have in place to assist knowledge exchange between areas?

c) Who is in charge of initial collaboration between researchers in your lab?

Score 1: Researchers do not support each other. There is no or only little knowledge exchange and collaboration.	Score 3: Senior staff try to use the right staff for the right job; researchers help each other (cross-project communication) but no supporting mechanisms are set in place to foster collaboration and knowledge exchange	Score 5: Staff recognize effective human resource deployment as a key issue and will go to some lengths to make it happen; enabling researchers to collaborate, reward intellectual challenging and transparent communication, cross-project and cross-lab-communication exists
Low	Degree of collaboration and knowledge exchange	High

Org - a) How many projects per researcher?	One project per researcher
	Multiple projects per researcher, no collaboration
	Multiple projects per researcher, with collaboration

Medical Research Survey

5) Performance Tracking

Tests whether performance is tracked using meaningful metrics and with appropriate regularity

Score:

1□ 2□ 3□ 4□ 5□

a) What kind of performance or quality indicators would you use to track the performance of your lab?
b) What do you measure to determine the success of a research project?
c) How frequent are these measured? Who can see it?

Score 1: Measures tracked do not indicate directly if overall objectives are met. Tracking is an ad-hoc process (certain processes aren't tracked at all)	Score 3: Most important performance or quality indicators are tracked; tracking is overseen by principal investigator only	Score 5: Performance indicators are regularly tracked and communicated to different groups, both formally and informally, to all staff using a range of visual management tools(e.g. Progress bars, milestones etc.)
Low	Degree of meaningful indicators and regularity	Hig

6) Performance Review

Tests whether performance is reviewed with appropriate frequency and indicators

Score:

1□ 2□ 3□ 4□ 5□

a) How do you review these performance indicators for your lab? …and for research projects?
b) Tell me about a recent meeting.
c) Who is involved in these meetings? Who gets to see the results of this review?
d) Probing: What tools/media do you use? (e.g. charts, timeline etc.)? (e.g. monthly report)
e) Do you have a follow up plan?

Score 1: Performance is reviewed infrequently or in an un-meaningful way (e.g. only the entire lab meets once in a while)	Score 3: Performance is reviewed periodically with both successes and failures identified; Results are communicated to everyone; There are lab meetings and "in-person" meetings	Score 5: Performance is continually reviewed (including meetings and reports), based on **different** progress/success indicators; All aspects are followed up to ensure continuous improvement; Results are communicated to all staff
Low	Degree of frequency and complexity of	High

7) Performance Dialogue

Tests the quality of review conversations

Score:

1□ 2□ 3□ 4□ 5□

a) How are these meetings structured? How is the agenda being determined?
b) Do you find you have generally enough information during the review?
c) Who is involved into the dialogue? (i.e. How much feedback comes from you vs. your researcher)
d) What type of feedback occurs in these meetings? For a given problem how would you identify the root cause?

Score 1: The right information for a constructive discussion is often not present; conversations focus overly on data that is not meaningful; **a clear agenda is not known and purpose is not explicitly stated**; next steps are not clearly defined. Only the PI speaks and discusses the projects.	Score 3: Review conversations are held with the appropriate data present (e.g. progress, next steps, challenges); objectives of meetings are clear to all participating and a clear agenda is present; The PI gives majority of feedback but everyone speaks up.	Score 5: Regular review/ performance conversations focus on problem solving and addressing root causes; purpose, agenda and follow-up steps are clear to all; meetings are an opportunity for constructive feedback and coaching and everyone is encouraged to speak his mind about the discussed topic.
Low	Degree of quality & participation of review	Hig

8) Consequence Management

Tests whether differing levels of performance of projects (not personal but plan/ process based) lead to different consequences

Score

1□ 2□ 3□ 4□ 5□

a) Let's say you've agreed to a follow up plan at one of your meetings, what would happen if the plan wasn't enacted upon?
b) How long is it between when a problem is identified to when it is solved? Can you give me a recent example?
c) How do you deal with repeated failures in a specific research project?

Score 1: Failure to achieve agreed objectives does not carry any consequences	Score 3: Failure to achieve agreed results is tolerated for a period before action is taken	Score 5: A failure to achieve agreed targets drives retraining in identified areas of weakness or moving individuals to where their skills are appropriate
Low	Severity and speed of consequences	High

Medical Research Survey

9) Types and Balance of Targets

Tests whether targets cover a sufficiently broad set of metrics and whether quantitative and qualitative targets are balanced

Score:

1□ 2□ 3□ 4□ 5□

a) What types of targets have you set for the research lab as a whole? (Ask for quantitative and qualitative goals)
b) Do you have intermediate targets/milestones?
c) How well do you think, do your researcher understand how the different targets are connected?

Score 1: Goals are exclusively focused on institutional targets and acquiring funding. (education of research fellows is not seen as a target)	Score 3: Goals are balanced set of targets (being especially academic quality and education). They are set by PI only and do not extend to all staff groups; interdependencies between goals are not well understood.	Score 5: Goals are a balanced set of targets covering three dimensions (academic quality, funding, education); interplay of all dimensions are well understood by senior and junior staff (PIs, Postdocs and PhD-Students).
Mono-dimensional	Differentiation of targets	Multi-dimensio

10) Interconnection of Targets

Tests whether targets are tied the organization's objectives and how well they cascade down the organization

Score:

1□ 2□ 3□ 4□ 5□

a) Do you have an overall goal/vision of your lab? How do you communicate this?
b) Are the projects linked to an overall goal or research question?
c) How are these goals cascaded down to all researchers?

Score 1: Goals do not cascade down the organization. No vision is in place and choice of research projects do not follow overall goal.	Score 3: Goals do cascade, but only to some staff groups (e.g. postdocs only). Only the PI sees the interconnections between the projects. An overall vision is not place or is not communicated.	Score 5: Goals increase in specificity as they cascade, ultimately defining individ expectations for all researchers. Each other sees the vision of the lab and how each research project plays a role in it. The vision is in place and gets strongly communicated.
Low	Degree of interconnection and cascading	H

11) Time Horizon of Targets

Tests whether the lab breaks down research questions into reasonable sub- experiments and has a '3 horizons' approach to planning and targets

Score:

1□ 2□ 3□ 4□ 5□

a) How do you go from broad research questions to smaller milestones?
b) How do your researcher prioritize their resources (time, effort) across the different projects?
c) What kind of time scale are you looking at with your research questions?

Score 1: Researchers have to coordinate on their own how to reach the research goals. They handle different projects without prioritization. The research projects are focused on short term goals (e.g. fast publication)	Score 3: Project is broken down into concrete milestones. Researcher prioritize their resources on their own but do not communicate that to the team. There are short and long term research goals which are not necessarily linked to each other.	Score 5: The research question in broken down into concrete experiments. The different resources are clearly prioritized into low to high priority and communicated to all labs staff. Short and long term research
Low	Time horizon and detail of planning	H

Org – b) Who sets the target and milestones for the research project?	Researcher
	Principle Investigator
	Both

Medical Research Survey

12) Target Stretch

Tests whether targets are based on a solid rationale and are appropriately difficult to achieve

Score:

1□ 2□ 3□ 4□ 5□

a) How are the targets perceived by the postdocs? Do they feel pushed by them?
b) On average, how often would you say that you meet your targets?
c) Do you feel that all research teams receive the same degree of difficulty, in terms of targets?

Score 1: Goals are **too easy** and only set by the researcher or the PI.	Score 3: In most areas, the principal investigator pushes for aggressive goals, but with little buy-in from researchers; there are a few sacred cows that are not held to the same standard	Score 5: Goals are genuinely demanding for all teams of the research laboratory and developed in **consultation** with principal investigator. Everyone has its own degree of difficulty.
Low	Degree of difficulty	High

13) Clarity and Comparability of Goals

Tests how easily understandable performance measures are and whether performance is openly communicated to staff

Score:

1□ 2□ 3□ 4□ 5□

a) Does anyone complain that the targets are too complex?
b) How do people know about their own performance compared to other people's performance?
c) When do you give feedback about personal performance? Do your researchers give each other feedback?

Score 1: Individual performance is not made public. Successes do not get celebrated and there is no opportunity for researchers to see how well they are doing compared to others.	Score 3: Performance measures are well defined and communicated; differences in performance visibility is in place but not gets foster by external mechanisms	Score 5: Performance measures are well defined, strongly communicated and reinforced at all reviews; performance is made public in a structured manner to induce competition
Low	Degree of open, clear communication of performance and comparability	Hig

14) Rewarding high performers

Tests whether there is a systematic approach to identifying good and bad performers and rewarding them proportionately

Score:

1□ 2□ 3□ 4□ 5□

a) Besides the paper and research itself, how do you reward very good researcher? (e.g. better space, going to conferences, mentoring).
b) How clear are people about who gets first and co-authorship once the paper is published? how their work will be recognized once the research is published

Score 1: People within the lab are rewarded equally irrespective of performance level.	Score 3: Very good researchers get rewards beyond publishing a good article (e.g. more mentoring or going to conferences). This is not done in a transparent or structured way.	Score 5: Clear performance accountability and rewards are set in place. (e.g. publish a research topic and you can go to a conference)
Low	Degree of differentiated rewards	High

15) Removing Poor Performers

Tests how well the organization is able to deal with underperformers

Score:1□ 2□ 3□ 4□ 5□

a) If you had an employee who could not do his job what would you do? Could you give me a recent example?
b) How long would underperformance be tolerated? (also ask for technicians and not only researchers)

Score 1: Poor performers are rarely removed from their positions	Score 3: Suspected poor performers stay in a position for a few months and are then moved into less important projects (i.e. contract will not be extended)	Score 5: We move poor performers out of the lab disregarding if they have own funding or not after a short period of time if they do not improve. There is a structural improvement plan in place.
Low	Severity and speed of consequences	High

16) Promoting High- Performers

Tests whether promotion is performance based and whether talent is developed within the organization

Score:

1□ 2□ 3□ 4□ 5□

a) How does your promotion system work?

b) How would you help him/her achieve his next position? Do you help everyone in the same way?

Score 1: People are promoted or referred to other institutions primarily upon the basis of tenure. The support for this is limited to a letter of reference.	Score 3: People are promoted upon the basis of performance. The Pi helps him to achieve the next position (e.g. letter of recommendation) but does not go all length or differentiates between people' performance actively	Score 5: We actively identify, develop and promote our top performers to other facilities. This is done by encouraging and mentoring them to find a new position
Low	Degree of support for the next position	High

17) Managing Talent

Tests what emphasis is out on overall talent management and continuous learning within the organization

Score:

1□ 2□ 3□ 4□ 5□

a) How do you show that attracting and developing talent is a top priority?

b) How do you support continuous learning of all your researchers?

Score 1: There is not much emphasis on educating: postdocs and PhDs beyond the necessaries.	Score 3:The principal investigator believes and communicates to postdocs that training talent is a priority. Some educational mechanisms are in place but most of the learning happens in an unstructured way.	Score 5: Continuous talent development is a top priority. There are processes in place (e.g. monthly lecture or journal clubs) to foster it. The PI pays great attention that everyone attends.
Low	Degree of educational processes	High

18) Attracting Talent/Recruiting process

Tests the strength of the employee value proposition

Score:

1□ 2□ 3□ 4□ 5□

a) What makes it distinctive to work at your research laboratory as opposed to your competitors?

b) What are you looking for in a "perfect candidate"?

c) How do you check on this? Can you guide me through the recruiting process?

Score 1: The selection process is one dimensional (e.g. only CV is checked) and only few applicants get screened.	Score 3: We are satisfied with application numbers. The selection process incorporates records, publication and referrals. Recruitment is not seen as a very important process in the lab.	Score 5: We can get highly selective in choosing researchers. Every application is carefully screened. The selection process is based on different factors and stages. interaction between candidate and the entire team is involved in it.
Low	Sophistication of recruitment process	Hig

Org - c) Utilization of Alumni:

Do you collaborate with your alumni? If so, how often and with how many of them?

There is no professional contact between PI and Alumni
There is regular collaboration but only in-frequently or with very few of the alumni
There is frequent collaboration with a high percentage of alumni. (exchange of resources, paper collaboration etc.)

Medical Research Survey

Org - d) When do they apply for funding? ("timing")	Demand driven (e.g. renewal of a grant)
	After a paper or data has been generated
	Continuously (independent of findings and ideas)

Org - e) Who decides on publishing the paper and what journal to?	Both	Org - f) Who is writing the paper?	Both
	Principle Investigator		Principle Investigator
	Researcher		Researcher

Self-reported success variables:

1. How much funding do you have per year?
    a. What is your success rate on funding applications?
    b. What percentage of your researcher has own funding (e.g. fellowships)?
2. How many paper does your lab publish per year on average?
3. What percentage of your alumni goes to academia?
4. How many patents you have?
5. How many serious applications do you receive per open position? (I mean all applications which are not spam applications)
6. To how many conferences do you go per year? __________
7. Do you review papers, if so how many per year? _________
8. Do you review grant applications (e.g. NIH), if so how many per year? ______

End of interview protocol.

Additional info:

- Operational Management: Q1-4
- Performance Management: Q5-8
- Target Management: Q9-13
- People Management: Q14-18
- Organizational structures: In-between questions “Org a-f)”