BIAS IN PUBLIC HEALTH RESEARCH: ETHICAL IMPLICATIONS AND OBJECTIVE ASSESSMENT TOOLS

A Dissertation

by

DANIEL VALDEZ JR.

Submitted to the Office of Graduate and Professional Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Chair of Committee,	Patricia Goodson
Committee Members,	Adam E. Barry
	Natasha Brison
	J. Timothy Lightfoot
Head of Department,	Melinda Sheffield-Moore

August 2018

Major Subject: Health Education

ABSTRACT

In an environment where one article is published every 20 seconds, we cannot be certain all studies are upheld to the same high quality standard. Thus, there is growing speculation that much of what is published today may contain embedded biases that detract from the quality of science. Though aware of bias in research, we are ill-equipped to address, identify and mitigate bias from published literature. Therefore, the purpose of this dissertation is to (1) explore the complexity and saliency of bias in published work via two domains: bias in numeric data (numeric bias), and bias embedded in language patterns (language bias) and (2) test technological tools intended to detect bias more objectively— namely the Cochrane Institute's GRADEPro, and topic modeling.

Numeric bias was defined as bias within number data and detected via the Cochrane Institute's GRADEPro software. To tout the effectiveness of using GRADEPro as a valid tool with which to detect number bias, this study used a heuristic example with currently published manuscripts on Pre-Exposure Prophylaxis (PrEP). Findings indicated, primarily, there were varying levels of evidence quality, ranging from Very High quality of evidence, to Very Low quality of evidence. Further, the efficacy of the medication in each study also varied by different extents.

Language bias was defined as bias within written language and identified more objectively via topic modeling. To demonstrate the effectiveness of topic modeling, I compared corpora of text data among three bias-inducing variables—time, funding

source and nation of origin. For each corpus, language patterns varied among the bias inducing variables, suggesting, among other considerations, bias inducing variables influence the direction of language despite testing the same hypothesis.

Overall, this dissertation sought to present tools outside of Public Health that could more objectively identify problematic issues within numeric and language data. For both types of bias, language and numeric, bias was identified and distilled in a more efficient and effective manner. Therefore, issues such as recurrent bias in Public Health should be addressed via these presented tools, as well as potential others, in the continued effort to uphold the integrity of science.

DEDICATION

This dissertation is dedicated to the women who instilled in me the importance of hard work, compassion, and a sense of humor— Grandma Lucy, Guela Chave, and my Mother, Isabel.

ACKNOWLEDGEMENTS

I would like to thank my committee chair, Dr. Patricia Goodson, for instilling in me the importance of critical thinking and her willingness to let me color outside the lines. I am forever grateful for the advice and encouragement of Dr. Adam Barry as well as his wisdom on how to navigate the complicated world of academia. Dr. Tim Lightfoot has always served as a source of inspiration, and his time on my project was invaluable and greatly appreciated. Finally, my thanks to Dr. Oi-Man Kwok and Natasha Brison, who, despite multiple time constraints, never hesitated to stop and address any questions I had along the way.

I would also like to extend my gratitude to various professors outside of Texas A&M and others who have since retired. Without the works of Eileen Gambrill or David Blei, I would not be enlightened in so many important topics that would eventually serve as the foundation, and primary method, for this dissertation. Most importantly, I am thankful for the work and personal correspondence with Dr. Bruce Thompson. Without his signature genius, dry as sandpaper wit, and genuine compassion for his students I would never appreciate statistics theory or the conceptual approach of statistics education, as it was taught to me.

I would be remiss to not extend a heartfelt hug to my beloved Physical Education Activity Program. From Melinda Grant— who took a chance and offered me a position as a graduate assistant— to Frank Thomas— the best boss anyone could hope for, and everyone else in between, PEAP has been a source of constant joy,

laughter and shenanigans and the best professional development on the planet. Special thanks to the backbone of PEAP, Amanda Nash and Sylvia Hubble, who both put up with my endless barrage of questions and never batted an eye when I locked myself out of the office. Lorinda & Dan Gomez, Mary-Beth Henthorne, Alyssa Locklear, Beth Netherland, Alex Pooley, Sara Safdari, K.B. Shea, Kristen Slagel, Dottidee Agnor, Teri Wenzel, Scott Wright, and the rest of the PEAP fam, your constant source of laughs and energy would get me through some of the toughest points in my program and I will never forget it.

Much of this journey is as personal as it is professional— and I am thankful for the best personal distractions any doctoral student could possibly hope for. Marc & Lisa Mendez, Jeffrey Glenn, Carol Rodriquez, Maura Casey, the Davila and Medina families, thank you for dragging me out of my house to go workout (yes, it can be too hot to exercise…and too cold….and rainy). To the few students I coached in Speech and Debate, Zoe Christianson in particular, I credit our work together as having sown the seed for mentorship. To my departmental family— Rahma Mkku, Jovanni Reyes, Leigh Szucs, Ming Li, Zack Jackson, and the Cubby Squad— thank you for your friendship even during my grouchiest of days. I would like to specifically thank Aditi C. Mukherji for keeping me sane, level headed, and laughing hysterically since our first encounter in the fourth grade twenty years ago. Cheers to our never-ending pursuit of establishing Choripan Inc, friend! And lastly, to Drew Pickett, whose unwavering support and encouragement motivated me, every day, to be better than the day before.

Above all else, I would like to acknowledge and thank my family and extended family for their never-ending patience with my never-ending schooling. Though I may live far away, I never stray away from my hometown roots. I would like to specifically acknowledge my family members who have since departed. To my Grandpa Eliseo Valdez, a staunch advocate for Latino worker rights, thank you for being a constant source of inspiration and light. I can year you yelling "Eso!" and "Ay Cabron!" with a biggest of belly laugh as I write this to you. Of course, to my late-Aunt Geraldine Agredano— the woman who was always first to comment on social media after an achievement, who was always the first at parties beaming with pride, who raised eight amazing children who are, themselves, raising amazing children of their own, and all with her signature sense of style, thank you for your warmth, sincerity, encouragement, and one of a kind laugh that could only be rivaled by none other than Mr. Valdez, himself.

CONTRIBUTORS AND FUNDING SOURCES

Graduate study was supported by a fellowship from Texas A&M University and supervised by a dissertation committee consisting of Dr. Patricia Goodson [advisor], Drs. Adam Barry and Natasha Brison of the Department of Health and Kinesiology, and Dr. Tim Lightfoot of the Department of Health and Kinesiology/Genetics. All work for the dissertation was completed independently by the student.

NOMENCLATURE

NHST	Null Hypothesis Significance Testing
GRADE	Grading of Recommendations and Assessment Developing and Evaluation
PrEP	Pre-Exposure Prophylaxis
P-HAART	Pediatric/Perinatal Highly Active Anti-Retroviral Therapy
TM	Topic Modeling
ES	Effect Size
HR	Hazard Ratio
CI	Confidence Interval

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER I

INTRODUCTION

Bias, a growing concern in 21st century academia, is defined here as any factor that sways credibility of scientific research outcomes (Delgado-Rodríguez & Llorca, 2004). Though the extent of bias within the scientific community is currently unknown, scholars in various fields are attempting to address underlying mechanisms potentially contributing to increased instances of biased, or misleading work. Inferential statistician Nuzzo (2014), for example, contends outdated reported practices, such as null-hypothesis significance testing, oversimplifies the complexity of hypotheses. Marketing professor Hubbard (2015) claims that due to corrupt research agendas, scholars also may be manipulating data deliberately to attain a favorable end. Neuroscientist Ioannidis (2005) ultimately concludes, based on these claims and other factor, that much of published research findings are partially inaccurate or false.

The drive to understand bias in scientific research stems, in recent times, from increased instances of manuscript retraction. Today, contrasted with the last decade, there has been a 44% increase in the amount of published research retracted by professional journals' editorial boards due to a host of concerns such as false peer review, failure to disclose funding source, conflicts of interest, data manipulation, and honest error (Nunberg, 2002; Cokol, Ozbay, & Rodriguez-Esteban, 2008; Wiles, 2012.). Therefore, as more instances of retraction occur, scholars have had a renewed interest into resolving such issues to uphold the credibility of science.

Many scholars (Barry et al., 2016; Nuzzo, 2014; Thompson, 2002) claim the current publishing system must address bias concerns by instituting reform. Thompson (1999), for instance, through the American Psychological Association, lobbied successfully in favor or requiring all manuscripts published in APA format include additional measures to denote study effectiveness, such as effect sizes. The Basic and Applied Social Psychology journal, in 2017, made a similar recommendation, contending significance tests hold no useful information and would be *optional* for future manuscript submissions (Trafimow & Marks, 2015).

Despite increasing interest in studying (and detecting) bias, and despite editorial recommendations for reform in reporting practices for professional journals, bias persists as a threat to scientific credibility. Most likely, it is because current means of identifying and detecting bias are limited, that the threat persists. Bias is, occasionally, a vague construct, but it is always difficult to identify (Bollen & Paxton, 1998). More importantly, because people differ in their beliefs and behaviors, correctly identifying something as biased or not biased is, often, opinion-based and subject to disagreement (Cook, 2014). Therefore, more stable, objective tools with which to identify and detect bias are warranted.

The purpose of this dissertation is two-fold: (1) expand the scope of research ethics to include language and numeric bias, and (2) showcase the utility of objective, technology driven tools – GRADEpro and Topic Modeling— with which to detect bias in numeric data and written language in published scientific reports The long-term goal of this project is to establish those tools—— as legitimate applications that should be

adopted by Public Health and other fields seeking to minimize bias within scientific reporting.

To explore these objectives most effectively, this dissertation utilizes a journal-article format. Each chapter will explore the following content: In Chapter 2 I will argue for expanding the scope of research ethics to include numeric and language bias as its components; components that should be considered as important as the rights of human and animal participants in research carried out in the 21st century. In Chapter 3 I demonstrate the utility and effectiveness of GRADEpro as a tool for detecting numeric bias in reported research. In Chapter 4 I present and demonstrate Topic Modeling as a language-bias detection tool and, Chapter 5 comprises conclusions and recommendations

Bias as a Component of Research Ethics

The first study, Chapter 2, seeks to frame bias in published work as a contemporary ethical concern. More importantly, this chapter also serves as the foundational backbone of the remaining two studies by defining terms, such as language bias and numeric bias, and introducing the theoretical perspective the remainder of the dissertation will follow. To introduce the concepts and frame bias as an ethical issue, I followed a multi-step process: (1) summarized, briefly, ethics and the importance of updating ethical codes of conduct, (2) defined what constitutes an ethical dilemma, and (3) logically mapped key points to determine if bias is, indeed, an ethical dilemma worthy of further attention.

Ethics is generally defined as the moral principles governing a person's behavior to distinguish right from wrong (Corlett, 2005). Throughout the history of ancient and modern thought, various forms of ethics emerged to create a fair and balanced environment that attempts to maximize the greater good (Byrne, 1988). Many fields working directly with human populations— such as Medicine, Law, and Public Health— have benefited greatly from codes of conduct intending to protect human, and animal, rights from unethical behaviors.

The drive to frame bias as a component of research ethics stems from multiple calls to update existing codes of conduct— Public Health, for example, has not updated its code of conduct since the early 2000's (Cohen, 2002). Historically, codes of conduct in applied fields such as Medicine and Public Health, have focuses on the protection of animal and human subjects participating in research. There are concerns, however, that factors other than the protection of animal and human rights could, themselves, pose an ethical threat to the credibility of research. Therefore, framing bias – or, more specifically, numeric and language-related biases -- as falling under the purview of research ethics validates its importance and allows for systematic examination of its occurrence and impact upon the current scientific enterprise.

Numeric Bias Identification with GRADEpro

Chapter 3 focuses specifically on numeric bias, or the bias embedded in numeric data and their statistical testing/reporting. Numeric bias is perhaps the most widely studied and commonly published type of bias in scientific studies. Calls for editorial reforms within professional scientific journals, for instance, are typically

supported by evidence generated through studies of *p*-value bias, or other biases related to statistical assumptions, data computation or results presentation (Ioannidis, 2005).

The purpose of Chapter 3 is to introduce and demonstrate the use of tools to (a) assist in detecting numeric bias, and (b) objectively evaluate the quality of evidence, vis-à-vis the potential for bias in the reporting of findings. To accomplish these tasks, I thoroughly discuss common data-related biases and their salience in published research. Further, as a heuristic example to test the utility of GRADEpro— one of the tools I propose for the detection of numeric bias— I will analyze clinical studies published on one of the most recent interventions fueling health policy debate.

Topic Modeling as a Language Bias Detection Tool

Language bias refers to observable or non-observable language structures that ultimately shape a misleading message (Bollen & Paxton, 1998). Such biases can either be sub-conscious (i.e. the agent is not aware he/she is inherently biased) or conscious (i.e. there is a deliberate attempt by the agent to mislead an audience (Gambrill, 2011). Though words-based language is one of science's main tools of communication (one could argue that mathematics – a number-based language – is, in fact, *the* language of choice for communicating science), studies of language bias are uncommon when compared to other types of biases, such as numeric bias (Zimbardo, 2010). Therefore, exploring language bias in greater detail can potentially give insight into how and why certain factors influence and shape scientific language, for promotion of specific research agendas.

The purpose of Chapter 4 , then, is to explore, in broad strokes, various bias-inducing factors that affect the language in research reports: (1) time, (2) funding source, and (3) nation of origin (Clifford, Barrowman, & Moher, 2002). More importantly, this chapter introduces Topic Modeling— a series of computer algorithms programmed to dissect large collections of text into latent topics (Blei, Ng, & Jordan, 2003)— as a novel tool with which to potentially detect nuances in published texts.

Specifically, in that chapter I examine whether Topic Modeling can correctly identify nuanced changes in language (due to time, funding source, or country of origin) within large collections of texts in various health-related fields. I explore the following questions in this chapter: (1) Can topic modeling identify changes in language over time in Ritalin research? (2) Can topic modeling identify nuances in federally- versus industry-funded studies on sugar consumption in the human diet, and (3) Can topic modeling identify differences between the United States and European Union language patterns for reporting on Pediatric/Perinatal Highly Active Anti-Retroviral Therapy (P-HAART)?

Overall, this dissertation seeks to create a well-rounded, conceptual understanding of bias and its presence in scientific research reporting. More importantly, this dissertation also seeks to highlight the utility of newer and more objective methodological tools for detecting and assessing potentially harmful biases in research. Results from this dissertation should contribute, at the very least, to the dialogue amongst scholars seeking reform for today's current scientific environment. In an era in which one article is published every twenty seconds (Bowman, 2014), the

scientific community must be ever-vigilant for the potential for bias that can, ultimately, damage the credibility of science and the public's well-being.

CHAPTER II

BIAS AS A COMPONENT OF RESEARCH ETHICS

With contemporary research issues such as experimental replicability (Ioannidis, 2005), false findings (Hubbard, 2016), and complications with the peer-review process (Henderson, 2010) emerging in science, there is renewed interest in the conceptual understanding of research ethics as it applies to scientific inquiry. Research ethics has been traditionally defined as, "the ethics of the planning, conduct, and reporting of research…that…should include protection of human and animal subjects," (Resources for Research Ethics Education, 2016).

While definitions such as the one above are well established, mainly due to strong theoretical support and ethical protocols most of them tie research ethics to human or animal rights. Today, however, newer ethical challenges highlight the contention that research is evolving and, thus, in need of protocols and definitions that address these challenges. Indeed, many groups and think-tanks are on the cusp of reconceptualizing research ethics to focus *both* on the dimension of the participant in scientific experiments (human or animal), and that of the researcher. For example, the Center for Research and Reform Education (CRRE) (2017) expanded its original definition (2000) to now state:

> not all researchers use human or animal subjects, nor are the ethical dimensions of research confined solely to the protection for research subjects. Other ethical challenges are rooted in many dimensions of research including the collection, use, and interpretation of data, [and] methods for reporting and

reviewing research plans for findings (retrieved from: http://archive.education.jhu.edu/research/crre/).

In its most recent iteration, the Declaration of Helsinki[1] (DOH) (2013) included an addendum with a similar position on the reconceptualization of research ethics:

> Researchers, authors, sponsors, editors, and publishers all have ethical obligations with regard to the publication and dissemination of the results of research. Researchers have a duty to make publicly available the results of their research on human subjects and are accountable for the completeness and accuracy of their reports. All parties should adhere to accepted guidelines for ethical reporting. Negative and inconclusive as well as positive results must be published or otherwise made publicly available. Sources of funding, institutional affiliations and conflicts of interest must be declared in the publication. Reports of research not in accordance with the principles in this Declaration should not be accepted for publication (p.2194).

Unfortunately, the drive to re-conceptualize research ethics stems from negative, rather than positive ideals. Driving many of these re-conceptualizing attempts is the growing influx of cross-disciplinary cases in which data appear manipulated,

---

[1] The Declaration of Helsinki, a preceding set of codes following the Nuremberg Code, is a set of ethical guidelines created by the World Medical Association in response to unethical human experimentation during WWII. Today, the declaration is widely viewed as the cornerstone of human research ethics (Moreno, 2017).

mal-explained, or intentionally/unintentionally altered. More importantly, the rate at which manuscripts are retracted across disciplines also raise concern" (Wiles, 2014).

The conceptual changes in the CRRE and the DOH's stances on research ethics, however, reflect a paradigm shift in research ethics. Rather than focusing almost entirely on animal and human protection, today, methodological decision-making is emerging as an important ethical concern that must be addressed.

The purpose of this chapter is to explore the theoretical underpinnings of the efforts to reconceptualize research ethics. To do so, I examine, in broad strokes, various methodological concerns that have surfaced in 21st century research. In effect, this analysis will highlight important reasons *why* research ethics should encompass both methodological decision- making alongside participants' rights. Equally important, this analysis will focus on elucidating the problems surrounding specific methodological issues in contemporary research (in the sciences and social sciences), namely: replicability of research findings and the factors affecting studies' replication such as manipulation of statistics, and researcher bias.

Concerns in Today's Research Climate

While science continues to, and will constantly face ongoing ethical debates regarding human subjects (and animals) who take part in scientific experiments (e.g. stem-cell research, cloning) it is also continually facing new ethical challenges related to the actual *practice* of science. In other words, the manner in which science is conducting itself in the 21st century is raising eyebrows among scientists and ethicists, above and beyond concerns with the well-being of research study participants.

The sheer volume of the scientific enterprise provides ample fodder for problems and difficulties. Consider, for instance, that current estimates suggest global scientific output doubles every nine years (Van Noorden, 2014). According to Munroe (2013) "a new [scientific] paper is now published roughly every 20 seconds" (p.59) and, in 2014, there were as many as 28,000 journals available for publication (many of which are considered "predatory", following unethical publication practices) (Larson, 2010). As the science industry continues to grow, more manuscripts are published than ever before, and more people seek doctorates (or research training degrees), today, than in the entire history of academia (Larson, Ghaffarazadegan & Xue, 2013).

The growing volume of information and/or terminal degrees awarded are not concerns in themselves. Many scholars have been worried, for several decades now, about the unscrupulous methodological practices that can *result* from over-publishing and diploma-dispensing. Among these problems are the manipulation of statistics, the inability to replicate research results, and researchers' biases. Rosnow & Rosenthal (1989), for example, expressed their concerns three decades ago regarding the manipulation of statistics, contending:

> We are concerned with the…strict logical consequences of statistical data analysis to shore up facts and inductive inferences. Despite the great range of procedures employed, there are some common problems of methodological spirit and methodological substance that although they have been addressed before, nevertheless endure. They are (a) the overreliance on dichotomous significance testing decisions, (b) the

> tendency to do many research studies in situations of low power, (c) the habit of defining the results of research in terms of significance levels, alone and (d) the overemphasis on original studies and single studies at the expense of replications (p.1276).

Nearly three decades after Rosnow and Rosenthal's assertion, little has changed to address the four methodological concerns they raise, apart from modest reforms from journal publishers 'encouraging' better reporting (American Psychological Association, 2016). Consequently, there is growing speculation, and incidences of uncovering suspect findings in published investigations.

Attributed, at least partly, to lack of methodological reform, scholars are also growing increasingly concerned that much of the work already published may not be replicable. Currently, more replication trials than ever are attempting to uncover flawed science. Consequently, the publishing community is experiencing unrivaled rates of retraction in major academic journals (Marcus & Oransky, 2014; Steen, Casadevall, & Fang, 2013) .

The pursuant question, then, becomes: what is driving 'poor' science and why? While many point to unintended error and oversight as major factors, others posit that perhaps implicit and explicit biases are to blame. For example, a researcher may be biased against certain groups, therefore, those groups are unintentionally omitted or ignored. Or, perhaps, the research is agenda-driven— meaning there is incentive to publish certain findings even if those findings do not adequately reflect the sample at hand.

Given the rise in instances related to the problems outlined above (for further details, see sections below) there is a renewed movement among scientists, researchers and scholars in various fields to have research ethics *also address and guide methodological decision-making, with the goal of upholding the credibility and integrity of science* (Gawande, 2016). To illustrate this need for expanding research ethics' scope, and to stand in support of groups that already have extended that scope, I will explore one current example and use its implications as a heuristic model to determine if methodological decision-making is, indeed, an ethical issue. In effect, this example will illustrate the need to expand the scope of research ethics to also include methodological practices for the continued improvement and renewal of science.

Replicability, Methodological Reporting, and Researcher Bias in Public Health Research

Complications with Result Replicability

The replicability crisis. Study replication is the hallmark of the scientific method (Open Science Collaboration, 2015). Science relies on replicable findings to uphold the generalizability of results reported in peer-reviewed manuscripts. While the process of replication has led to numerous advances in science, applied fields such as Psychology, Medicine, and Public Health are, today— according to many concerned researchers—at the cusp of a 'replicability crisis' (Diener & Biswas-Diencer, 2016).

The replicability crisis, simply defined, is the increased difficulty in duplicating or reporting a particular study's findings, even under identical study conditions (John & Loewentein, 2012; Cook, 2014). Feilden (2017) contends, for instance, this growing

inability to replicate poses serious implications that could lead to lasting, negative consequences including: (1) potentially false, published information, (2) decline in quality of academic work, and (3) public mistrust of published data.

To emphasize the inherent difficulty of replicating findings, social scientist Brian Nosek undertook an ambitious effort in 2011 to reproduce findings from 100 studies published in top-tier journals in the field of Psychology (Open Science Collaboration, 2015). As he anticipated, much of the findings were problematic. Out of the 100 test studies, less than half (n=39) reproduced findings identical to the parent study.

After completing his investigation, Nosek effectively listed several factors explaining the inability to replicate, such as the difficulty in quantifying a behavior, research error, and change of variables across time, among others. Overall, however, Nosek's conclusion was that even in the best cases of well-documented and 'clean' research, replication is difficult and hard to accomplish.

On an even larger scale, a group of 270 authors led by Aarts et al. (2011) published an article in Open Science Collaboration discussing how well 100 high-profile experimental findings in psychological sciences were replicated exactly as the original parent investigations. Similar to Nosek's (2015) assessment, only one-third to approximately one-half of studies were reproducible at any level. Aarts and colleagues asserted:

> …collectively these results offer a clear conclusion: A large portion of replications produced weaker evidence for the original findings despite using

> materials provided by the original authors, review in advance for methodological fidelity, and high statistical power to detect the original effect sizes. Moreover, correlational evidence is consistent with the conclusion that variation in the strength of initial evidence (such as original *p* value) was more predictive of replication success than variation in the characteristics of the teams conducting the research, such as experience and expertise (2012, p.aac4716-6).

The investigations pioneered by Nosek and the Open Science Replication Collaboration provided important evidence for the current difficulties with replicability in science and its varied contributing factors. Though useful, both studies operated under the assumption that replicating findings was difficult because of (1) accidental error, and (2) external stimuli beyond the control of either the original research team or those replicating the findings.

Inherently, valid scientific replication is difficult—especially in fields that deal with unpredictable variables such as human behavior (Artino, 2015). In replication studies among both the hard sciences and the social sciences, the goal is to produce findings identical to those in the parent investigation. However, in replications of social science studies, results tend to be not only marginally different, but also smaller in terms of effect sizes (measures of experimental effectiveness) due to the complexity of human behavior and other issues related to a particular sample (Bohannan, 2015). Beyond problems within the data, moreover, other external factors may also contribute to unintended changes in a study's environment, which impact the final results (e.g.

timing, location, skills of the research team, among others). Thus, due to the complexity of human behavior, even the best replication attempts are still prone to achieving results that may not reflect the original investigation's outcomes.

There is no replicability crisis. Not everyone agrees, however, with the notion there might be a replicability crisis in the social sciences (in particular). Another growing body of scholars has actively pushed away recommendations and suggestions for editorial reform proposed by institutional and editorial boards such as AERA and APA, claiming concerns are exaggerated. Pashler & Harris (2012) contend that not only does the crisis *not* exist, but "…these arguments that we have heard most often from scientists who see the current outpouring of concern over replicability [are] greatly overblown" (p.531). Specifically, detractors of the replicability crisis argue: (1) risk of Type 1 and Type 2 error are known conceptual flaws of *p*-values, however, researchers mitigate this issue by setting the threshold level for alpha at a tolerably low level of 5%, which allows Type 1 error to go unnoticed, and (2) while replication trials are rarely conducted, "researchers frequently attempt (and publish) conceptual replications, which are more effective than direct replications for assessing the reality and importance of findings" (p.533), and (3) science will self-correct and false findings will eventually be discarded.

Pashler & Wagenmakers (2012) explain journal editors' decisions to enact strategies to enhance replicability and to enforce policies promoting effect-size reporting, alongside increased calls for editorial vigilance, are nothing more than "a crisis of confidence" (p. 530). Pashlet & Wagenmakers further argue, such efforts

directly breach intellectual freedom and foster unwarranted policing of scholarly work. Consequently, an increased surveillance of academic work can destroy the impetus to produce innovative and groundbreaking science, claims Healy (2007).

Following a similar logic, Stroebe and Strack (2014) view 'exact' replications as unfair and impossible to accomplish in the social sciences. In effect, a replicability crisis cannot exist because the current standards are too high for social scientists. In the natural sciences, such as Chemistry or Physics, results from studies are easily replicated because the laws of nature are more stable. When dealing with human populations, as Stroebe and Strack (2014) contend, variables are more unpredictable, and a host of factors contributes to potential changes in the landscape of a study, ranging from characteristics of the sample to level of power selected for statistical analyses.

For example, studies testing the laws of physics, in theory, can easily replicate regardless of study conditions due to nature's stability. However, even the most popular and widely used instruments are subject to unpredictable human behavior, wherein factors such as mood, time of day, or trauma can alter the answers between sequential iterations. Thus, contends Peng (2015) instead of focusing on *exact* result replicability – researchers within the social and applied sciences should, instead, be concerned with *conceptual*[2] replicability.

---

[2] Conceptual replicability refers to applying the same theories and protocols in subsequent investigations, but ignores score reliability due to the unpredictability of human behavior (Pashler & Wagenmakers, 2012).

In sum, proponents of the 'no crisis' stance (in the replicability debate) claim there may be issues with replication, however, calling these issues a 'crisis' is inappropriate because (a) one cannot exactly replicate human behavior, and (2) science will self-correct by tried and true measures, such as blind peer review. Due to the complexity and variability of social sciences research, editors and researchers should allow science to stay as is. Any effort to increase vigilance or impose stronger editorial oversight will do nothing more than hamper the freedom of academia for no valid reason, claim the replicability crisis critics (Earm & Trafimow, 2015).

Replicability as an Ethical Issue

In effect, the replicability crisis is a complex and divisive issue among scholars. Regardless of the opinion one holds, however, complications with replication have both influenced and divided scientists on how research *should* be conducted. However, regardless of one's stance on the issue, concern over replicability begs an important question: are contemporary issues of result replicability and research studies' underlying methodological decision-making simply a squabble amongst intellectuals, or are they emblematic of more important ethical issues, warranting further assessment?

From what is known about ethics[3], applied ethics[4], and ethical frameworks (see Appendix A), there are multiple criteria distinguishing common problems from ethical ones. According to Braunack-Mayer's (2001) guidelines on what constitutes ethical dilemmas in medicine, there are two steps to determine if a given situation constitutes an ethical issue. First, one must determine if the situation is even a problem by evaluating it against two tenets: (1) agents must have a choice to make, and (2) there must be many potential outcomes stemming from the eventual decision.

Once classified as a problem, the situation can *only* be considered an ethical issue if the decision-making process is affected and influenced, by morality, values, and beliefs (Resnik, n.d.). In other words, when making a choice between one or more options, agents must carefully evaluate and choose between what is perceived as morally right and morally wrong. It is morality that uniquely distinguishes ethical issues from common problems (Mbidde, 1998). Unlike many common problems, which have definitive answers, ethical issues rarely have a perfect solution because moral judgement varies among agents (Braunack-Mayer, 2001). In other words,

---

[3] Ethics is broadly defined as a normative evaluation used to distinguish right from wrong based on varying dimensions, such as: morality, justice, obligation, and righteousness (Braunack-Mayer, 2001).

[4] Fields in the applied sciences, such as health promotion, Public Health, and medicine, among others, traditionally have relied on classical ethical thought to guide complex decisions. However, they often require a more "hands-on" or practical framework due to the nature of their practice (Resnik, n.d.).

regardless of the attempt to arrive at a fair and equitable solution to all parties, at some level, some agents stand to lose something.

According to Braunack-Mayer's listed criteria, issues with scientific replication constitute, at the very least, a *problem* in research (i.e., agents have choices and there are multiple potential outcomes for each choice). However, concerns with replicability also have moral implications, indicating the replicability crisis *is*, indeed, an ethical problem. First, because producing credible science is the expectation of all agents, there is an implied moral principle that it is *good*, or ethically appropriate, to produce quality investigations and *wrong* (unethical) to produce questionable science. In other words, the choice to make science replicable is guided by morality and exercising this choice becomes, therefore, an ethics-guided behavior.

Additionally (and, perhaps, more importantly), replication matters also are guided by morality when involving the lay public. If a scientist knowingly produces questionable or inaccurate research, there are moral implications when the general public is a consumer of that research. In the recent retraction by Springer of several Oncology research reports[5], for example, there are serious implications with regard to treatment of a patient's illness.

If by employing Braunack-Mayer's (2012) criteria we agree methodological practices affecting research replication can have significant ethical implications for

---

[5] In 2017, Springer Publishing Group retracted over 100 manuscripts, the largest instance of retraction in the history of science, due to allegations of false peer review. All manuscripts were published in Tumor Biology, often considered one of Springer's most prestigious journals (Stigbrand, 2017).

both scientists and the lay public, what, then, are the steps for addressing these practices and implications? Hanson (2014) contends that in professional and business ethics (both considered applied fields) the most important action one must take to address a dilemma is to understand and properly identify the various factors causing the dilemma, then apply field-specific frameworks as guides to navigate the ethical issue.

Unfortunately, many current frameworks for research ethics in the health sciences (our main area of concern) are primarily centered on the protection of human and animal subjects. For example, bioethics, the ethical framework guiding Medicine and, to an extent, Public Health, has numerous codes and strategies directly addressing human subjects (e.g. informed consent, and Institutional Review Boards). Thus, in the contemporary world of health-focused research, methodological concerns (or factors affecting replicability of findings) are studied from an ethics perspective far less frequently than concerns related to human or animal participation in research.

Fortunately, however, alongside previously mentioned groups and think-tanks, several national organizations such as the National Institutes of Health (NIH), the National Science Foundation (NSF), and the Food and Drug Administration (FDA), among others, are now overemphasizing the importance of quality standards for research. On its research ethics webpage, the National Institute of Environmental Health (NIEH) recognizes the complexity of the research process and the need to behave ethically on *multiple* fronts in research:

> Although codes, policies, and principles are very important and useful, like any set of rules, they do not cover every situation, they often conflict, and they require considerable interpretation. It is therefore important for researchers to learn how to interpret, assess, and apply various research rules and how to make decisions and to act ethically in various situations. The vast majority of decisions [need] straightforward application of ethical rules. (Resnik, 2015, retrieved from https://www.niehs.nih.gov/research/resources/bioethics/whatis/index.cfm)

The acknowledgement by the NIEH that research is complicated and requires ethical parameters is emblematic of the growing movement in which methodological decision-making is viewed as equally important as patient/animal rights. Furthermore, because the NIEH is not the only major federal agency to voice similar concerns (e.g. NIH, NSF, and FDA have done so, also), these apprehensions further legitimize concerns over replication of investigations. Therefore, I briefly outline and discuss below several factors affecting replicability, in the continued bid to frame methodological decision-making as an ethical issue.

Factors Affecting Replicability

Munafo & colleagues (2017) contend there are several factors that hamper replicability and, in and of themselves, contribute to the greater ethical challenges embedded in methodological decision- making. The discussion below focuses on the two most problematic factors : (1) manipulation of statistics reporting (accidental or intentional), and (2) underlying researcher bias (intentional or non-intentional). Both

issues, by nature, are extremely complex, and a full treatment of each is beyond the scope of this paper. However, examining specifically how these factors fuel broader ethical challenges within methodological practices is necessary to re-conceptualize and expand the scope of research ethics, as proposed earlier.

Manipulation of Statistics Reporting

At its most primary level**,** the field of statistics relies on decision-making (Thompson, 2006). In any given investigation, the researcher must carefully select the appropriate methodological analysis tool with which to address the research question. In public health, researchers primarily rely on the General Linear Model (GLM) to address sample-and -level questions via correlation-type analyses (Pearson *r*, t-Test, ANOVA, regression, among others) (Hayat, 2017).

While the logic of the GLM is sound due to its simplicity[6], the model is not immune to common errors that affect presentation of results. Simply stated, all analyses in the GLM abide by the same logic, mathematical calculations, and assumptions regardless of the complexity of the analysis (Field, 2014). Therefore, contends Hubbard (2016), given the simplicity of the general linear modeling, over the years researchers have developed misconceptions that harm the validity of findings. As Ioannidis (2005) warns, if misconceptions within statistical frameworks go unchecked, the majority of published findings could be false.

---

[6] The Logic of the GLM assumes all analyses are all special cases of each other. In other words, these analyses abide by the same logic/assumptions and should be used "as a unifying conceptual framework for teaching statistics and psychometric theory" (Thompson, 2015, p. 30).

Some scholars in the field of health promotion and behavior (Barry and colleagues, 2016) have conducted investigations to uncover patterns with reporting practices due to common misconceptions of the GLM. Barry et. al., along with other scholars across multiple disciplines, uncovered patterns related to statistical representations of data that warrant concern, such as: (1) overreliance and misuse of null hypothesis significance tests (Nuzzo, 2014, 2014a; Head, 2015); (2) underreporting of confidence intervals and effect sizes (Fan, 2001; Vacha-Haase & Thompson, 2004; Thompson, 2007; Barry et al., 2016; Barry, Valdez, Szucs, Reyes & Goodson (in press); and (3) misunderstandings of common psychometric properties (Pickett, Valdez & Barry, 2017).

Overreliance and misuse of null-hypothesis significance tests. A statistical mainstay in Public Health, null-hypothesis significance tests (NHST) are defined, in simple terms, as the probability the result of a tested hypothesis did not occur by chance (Field, 2017). Despite its common use in academia, however, criticisms of NHST, are almost as old as the method itself (Boring, 1919). Thompson (2002a) elaborates:

> Statistical significance estimates the probability ($p_{CALCULATED}$) of sample results deviating as much or more than do the actual sample results from those specified by the null hypothesis for the population, given the sample size. In other words, these tests do *not* evaluate the probability that sample results describe the population…Instead, the tests assume [incorrectly] that the null exactly describes the population and then test the sample's probability…*This*

logic is so convoluted that, as empirical studies confirm, many of the users of statistical tests indeed do not understand what these tests actually do.

(Thompson, 2002a p. 65, comment and emphasis added).

Dubious logic aside, *p*-values remain researchers' primary and preferred statistic for reporting results across academic disciplines. The popularity of the *p*-value likely stems from the simple, dichotomous nature of its interpretation: specifically, any probability greater than $p=.05$ renders a finding non-significant (Nuzzo, 2014). In other words, by relying on a simplistic 'yes' or 'no' framework, a *p*-value can very easily and simply inform whether a result did or did not occur by chance[7] (Wasserstein & Lazar, 2016). Apart from assigning statistical significance, however, the *p*-value serves no other informational purpose.

Due to the limited information provided by a *p*-value statistic, many scholars worry about how the practice of reporting *p*-values affects the quality of a study. For example, it is common knowledge that *p*-values are bound by sample size (e.g. an increase in sample size *n* leads to a smaller *p*-value) (Sullivan & Feinn, 2012). Therefore, if amending a non-significant finding is as simple as increasing the sample's size, then how can we be certain some findings have not been, "p-hacked….to scientific glory? (Aschwanden & King, 2015, retrieved from: https://fivethirtyeight.com/features/science-isnt-broken/#part1)

---

[7] A full discussion of the American Statistical Association's position on *p*-values can be found at: http://amstat.tandfonline.com/doi/abs/10.1080/00031305.2016.1154108#.WpHGbKjwaUk

Despite providing relatively little information about the hypothesis, *p*-values remain the mainstay in Public Health, Health Education, and Health Promotion—primarily due to their ease of interpretation (Nuzzo, 2014). Furthermore, the simplicity in which one can attain a significant *p*-value has also fueled a bias against non-significant findings (Levine, Asada & Carpenter, 2009). This bias, in which non-significant *p*-values are deemed non-important, has two important implications across the social sciences: (1) researchers may be selectively reporting their findings and not a true overview of their results, and (2) scientists are potentially manipulating data to attain a significant finding (Aschwanden & King, 2015). Such issues with *p*-values are the primary reason Thompson (2003) and Field (2007) contend one should always report other metrics, such as confidence intervals and effect sizes, to better understand the underlying meaning and relative importance of significant and non-significant findings.

Under-reporting of confidence intervals and effect sizes. Confidence intervals (CIs) and variance-accounted-for effect sizes (ES) have frequently been cited as plausible alternatives to significance testing (Fan & Thompson, 2001; Lee, 2016). Yet, these alternatives are often under-reported and under-utilized, further obfuscating research findings and their replicability (intentionally or not).

CIs are used to help generalize the findings from a sample to a population (Thompson, 2006). More importantly, they also articulate the precision of a mean score/value, otherwise known as a point estimate. To calculate the statistic, all experiments must first be concluded. Upon successful completion of experiments,

researchers will set a series of parameters to capture a hypothetical population parameter from the experiment's sample (Savage, 1972). First, they set a confidence level with which to test the hypothetical parameter (e.g. traditionally set at 95%, or 1-α – where α = .05) (Glover & Mitchell, 2015). The confidence level informs us that, in 95% of cases, hypothesized means will land within a certain range from its original point estimate, or initial value (e.g. upper and lower bound limits). The resulting upper and lower bounds limit the range of a point estimate (Field, 2017).

For example, let us imagine a group of researchers conducting experiments on a medication's effectiveness. Researchers found that, on average, the medication was shown to be 90% effective within the tested sample. To determine the strength of the point estimate, the research team calculated a 95% confidence interval with upper and lower bound limits. If the confidence interval around the 90% point estimate was, hypothetically, 85%-95%, we can infer that in 95% of cases in our hypothetical population distribution, the range of effectiveness would fall between 85 and 95%. Conversely, if the researchers had arrived at the same point estimate (90%) but had wide margins in their confidence interval (e.g. if the CI was 21%-90%) we would infer that, if 100 samples were randomly selected from that population, in 95 of them the medication could exhibit varying levels of effectiveness, between 21% and 90% effectiveness—a finding that might change the view of the medication's utility.

While there are various types of effect sizes, the majority determine the magnitude of a treatment's effect. In broad strokes, an effect size is generally viewed as the percent of variance explained by a given statistical model, or how well an

independent variable explains the variability in a dependent variable. All GLM analyses have a corresponding effect size. Within the GLM framework the effect sizes are analogous to each other.[8] Because all GLM analyses share common traits, it is easier to compare effects across studies using ES, regardless of the analysis used.

Updated reporting recommendations for the journal American Psychologist strongly encourage (Zimbardo, 2002) researchers to report, along with *p*-values, some measure of the magnitude of effectiveness. The purpose of providing an effect size is to furnish that measure of the magnitude of the effectiveness associated with a significant *p*-value. For example, if a study's finding hypothetically attained a significant *p*-value ($p<.05$), but had an ES of .12 (or 12%), how important would that finding be? In this specific example, we might not place much faith on that finding because, while significant, the size of the effect is quite small (Sullivan & Feinn, 2012).

Though CIs and ES provide invaluable information beyond significance tests alone, they are reported much less frequently than significance tests. Barry and colleagues (2016), for instance, found that within 1,245 studies in five flagship Health Education and Promotion Journals less than half (47.9 %) reported an effect size, despite strong recommendations by the APA to do so. Unfortunately, the trend to underreport in Health Education and Behavior is not unique, as the majority of academic disciplines in the social sciences face similar difficulties (Trusty, Thompson, Petrocelli, 2004; Sullivan & Feinn, 2012).

---

[8] Pearson Correlation (r), t-Test (Cohen's *d)*, Analysis of Variance (Eta^2), Regression, (R^2)

Two implications stem from underreporting of CIs and ES. First, without the valuable information provided by these metrics, one cannot be certain about the validity or the practical importance of research findings. Second, without a hypothesized range from a hypothetical population parameter, or information about the percent of variance explained by the statistical model, replication studies are less effective. In other words, replications become difficult when there is not enough information available to compare to the replicated findings and determine, with any degree of confidence, whether the replicated results are identical to the original ones (Lackens, 2013).

Misunderstandings of psychometric properties. Over-reporting of *p*-values and under-reporting of CI and ES relate to the presentation of a study's results. The psychometric properties of data collected through survey-based research are equally important and problematic. None, however, cause as much confusion as *unique score reliability* (Vacha-Haase, 2002).

The score reliability statistic determines if answers given to survey questions follow similar or divergent patterns within the group (sample) responding to the survey (Thompson, 2000). Each group of respondents will exhibit its own distribution-of-responses pattern, rendering reliability a characteristic of the *data from a specific sample*, and not of the instrument (or the questions/survey instrument). Unfortunately, it is all too common for researchers to forgo reporting the score reliability coefficients of their own data and, instead, report a coefficient appearing in an accompanying testing manual or in previously published studies using the same survey questions

(Thompson, 2002b). Such actions are incorrect, as 100 samples will almost always yield 100 different reliability estimates (Vacha-Hasse, 2001; Tani, Logan, Woodall & Thompson, 2002).

To confirm the notion that reliability is a function of the sample, Pickett, Valdez & Barry (2017) examined the alpha[9] reliability coefficient for measures of delinquency among teenagers from 8th, 10th, and 12th grade sub-samples across a 40-year span from Monitoring the Future Survey. As expected, the alpha coefficients exhibited small fluctuations with every yearly iteration of the survey. One important distinction, however, was the 12th grade group, who had significantly lower levels of reliability over time, compared to their 8th and 10th grade counterparts. This analysis begs the question, if 'age' was a strong enough variable to influence groups' response patterns on a survey, might other variables such as 'race' or 'sexual orientation' affect score reliability of different groups?

Thompson (2002b) contends failure to understand score reliability, and the continued practice of reporting coefficients from prior studies and/or testing manuals, holds two implications for research: "poor score reliability may compromise…score validity [and] may compromise the ability of a study to yield noteworthy effects" (p.5). In other words, by reporting coefficients from previously-conducted studies, or failing

---

[9] While there are many reliability coefficients, the most commonly applied in Public Health and Health Promotion and Behavior is Cronbach's Alpha, which is a measure of how consistent the responses to a set of items are. If consistent, the responses are understood to be measuring a common, latent variable (in other words, the survey items are measuring the variable they intend to measure). (Thompson, 2002)

to report alpha coefficients all together, we cannot be certain the data, themselves, are valid. This leads to important consequences when attempting to replicate a study, given that many of the complications associated with the psychometric properties of the data will not emerge (or be identified) until the study is replicated.

In response to concerns with unique score reliability, the APA Special Task Force (1999) suggests, "Interpreting the size of observed effects requires an assessment of the reliability of scores" (p.596). However, this recommendation, in concert with not relying on *p*-values as the sole criterion for an effective result, as well as encouraging the inclusion of more CIs and ES, are often ignored.

Consequently, the inability to replicate continues to be fueled by these underlying issues that are known, but relatively unaddressed (particularly in health promotion and behavior research) Separately, these three issues— overreliance and misuse of NHST, underreporting of CIs and ES, and misunderstandings of psychometric properties— seem easily remedied through increased emphasis in statistics education. They are, however, issues coalesced into a much larger problem in which the quick proliferation of manuscripts is more important than study quality. As mentioned, concerns with methodological reporting are not new. Today, however, these concerns represent an emergent ethics consideration in the Health Sciences.

Methodological Reporting and Bias

In scientific investigations, bias is best defined as any deviation from the truth in data collection, analysis, interpretation, and publication resulting in the presentation

of false conclusions as truth (Simundic, 2013). Smith and Noble (2014) elaborate that understanding and studying research bias in multiple contexts is important because:

> …first, bias exists in *all* research, across research designs and is *difficult* to eliminate; second, bias can occur at *each stage of* the research process; third, bias impacts on the validity and reliability of study findings and misinterpretation of data can have important consequences for practice (p.2, emphases added).

Most scholars and consumers of scientific information understand that, at some level, personal or professional decisions may be driven by implicit biases that can potentially affect research. There are, therefore, multiple protocols, procedures, and taxonomies in place intended to mitigate bias in scientific investigations. Though these protocols are intended to serve an important purpose (i.e. bias mitigation), they are, as with research reporting suggestions, merely recommendations that are oftentimes ignored.

In research practice, perhaps the most sweeping taxonomy of bias is the work of Sackett (1979) who proposes nine different types of biases that can occur at respondent and investigator levels[10]. Updated in 2014 by Sarniak (to expand upon the effect of bias in research), these taxonomies hinge on the assumption that bias can be

[10] The types of biases in Sackett's taxonomy include: (1) Acquiescence bias, (2) Social Desirability bias, (3) Habituation bias, (4) Sponsor Bias, (5) Confirmation bias, (6) Culture bias, (7) Wording bias, (8) Question Order bias, and (9) The Halo Effect bias

identified and sorted into fixed groups. In reality, however, bias is far more complex and nuanced than these taxonomies would suggest.

Hammersly & Gomm (1997) contend that apart from pre-existing taxonomies on bias, the study of bias itself "has been given relatively little attention in the methodological" sense (p.68). More importantly, due to its complex nature, there is still a lot to learn regarding the mechanics of *how*, practically and conceptually, bias affects the quality of research.

Herein, I will argue, specifically, that bias affects the quality of research, and I will support this argument by examining bias as a complex phenomenon rooted in human behavior. This approach to viewing bias is paramount to: (a) mitigate bias more effectively, and (b) contribute to the greater debate of expanding the scope of research ethics. Effectively understanding how bias impacts research serves to potentially re-evaluate our approach to studying bias in academic work. More importantly, having a fully realized and conceptual understanding of bias is key to understanding the need for a more comprehensive approach to research ethics.

Intentional and Unintentional Biases. Staats & Patton (2014) posit that bias is inherently difficult to capture because most biases are implicit. In other words, biases are almost always internalized, and the manifestations of bias can occur at a subconscious (unintentional) rather than conscious (intentional) levels. Because bias is implicit and difficult to identify, there is frequent debate regarding what type of bias (intentional or unintentional) is more common in research. While both types of biases warrant concern among the academic community, successfully distinguishing between

unintentional or intentional bias is key in determining how to mitigate the interference of bias in published scientific work.

Ransohoff (2005) contends that unintentional bias may not be malicious, necessarily, but may, rather, stem from tradition or from preferred practices. In statistics, for example, mean imputation is a proven liability for handling missing data, especially when the method is not disclosed in the published manuscript (Donders, van der Heijden, Stijnen & Moons, 2006). Some scholars, however, may continue to employ this method without being fully informed about some of its inherent flaws. Mean imputation, then, can result in one, or many publications that are biased because the data are neither accurate nor truly reflect the sample (Allison, 2002).

In opposition to the notion that most biases are unintentional stand Gupta and associates, (2009), McArdle, (2011), and Newsome, (2015) who believe bias is intentional, purposeful and agenda-driven. In their view, many instances of biased work stem from a desire to produce outcomes that achieve some favorable end. In our mean imputation example, while one investigator may impute the means in his/her dataset due to lack of knowledge of the inherent flaws in the procedure, another researcher may use mean imputation to treat missing data knowing full well this choice can result in inflated Type-1 error and change the outcome of a hypothesis (Wicklin, 2017).

The scholars who believe in, and are currently addressing, the replicability crisis in science fall into this latter line of thinking— according to them, published findings are not replicable due to biases shaping (a) data collection, (b) statistical

calculations, (c) result reporting, or (d) peer-reviewing -- biases that represent deliberate attempts to boost metrics and/or handle external pressures and conflicts of interest. Conversely, those who contend the only 'crisis' in academia is a crisis of confidence stand with scholars who believe bias is a non-issue and, most likely, unintentional.

Regardless of motivations, both unintentional and intentional bias can be equally problematic. Though intentional bias may be associated with hidden political agendas (McGarity & Wagner, 2008), being uninformed about issues regarding bias could pose equal, or greater harm, especially in academic fields studying vulnerable populations (Ransohoff, 2005). For example, if a researcher is unintentionally biased against underrepresented minority groups but studies health disparities, to what extent does the work of that investigator truly reflect the population? In this example, the researcher may not be aware he/she is creating a biased environment in which groups may be treated differently (Burgess, Van Ryn, Dovidio & Saha, 2007). Conversely, however, manipulating data to gain a favorable hypothesis, whether through deleting/adding cases, applying inappropriate weights, or altering sample size to attain a favorable end pose serious implications for the credibility of science and its institutions (Simundic, 2013). In both instances, the biases built into the investigations can have damaging effects for those who participate in, or consume, the resulting findings.

Moreover, both situations outlined above are examples of poor judgement and lack of personal/professional reflection. Unintentional biases indicate poor personal

judgement that consequently affects the world around each respective agent. Intentional biases, such as data manipulation, are contributing to the increasing mistrust of academic institutions as well as to the increasing concerns of growing academic greed and flaws in the peer-review system (March, Jayasinghe & Bond, 2008; Bohannon, 2013). Yet, despite knowing, at least on some levels, of the existence of bias, calls or suggestions to improve bias detection are scant in the methodological literature. Yet, without working toward improved bias detection, problems with bias in academic work will remain ongoing (Sica, 2006).

When attempting to detect, or identify, biases in research we are currently left with one approach: making moral judgement calls to determine if work *is* or *is not* biased (Chan & Altman, 2005). While this verdict-driven approach to handling bias has proven somewhat effective and successful in identifying problematic patterns in published work (e.g. Springers' retraction of 107 articles due to fake peer review), it fails to address the complexity and multi-dimensional layers of bias and bias inducing factors, alongside failing to determine *how* those factors contribute to decision-making. Thus, by relying on a judgement/verdict approach (e.g. this *is* versus this is *not* biased) to identify bias we cannot definitively stake a claim that something is or is not biased without making challengeable assumptions about the accused individual/group' decisions.

Scholars have long been critical of current approaches to detect bias, citing the need for improved research practices and peer review via more sophisticated dynamics (Stoker, 1995; Smith, 1997; Casadevall, Fang, & Morrison 2009). I argue in support of

those scholars who contend bias detection should be as sophisticated as bias, itself. Without some type of reform, or at least improved capabilities to detect it, bias -- intentional or not -- will continue to limit trust in and continual development of the scientific enterprise.

The subjectivity of bias. The most salient reason verdict assigning is ineffective at identifying and deterring bias stems from one common, inescapable element: bias identification is subjective (Bollen & Paxton, 1998). Simply put, what may seem biased, or exhibit indications of bias, to one researcher may be perfectly acceptable for the other due to a conceptual gray area in which personal and professional opinions influence key positions on important issues.

In qualitative inquiry, for example, subjectivity is at the forefront of the investigation due to potential bias when interacting with participants, coding transcriptions, and interpreting results (Ahktar-Danesh & Bowman, 2008). Consequently, qualitative scholars must be constantly vigilant that their own personal or professional beliefs and factors do not play a role in shaping the outcomes of interview-based investigations (Collier & Mahoney, 1996). However, even if the qualitative scholar was thorough, careful, and completely stringent in guaranteeing his/her work is free of bias, the investigator is still not immune to critics seeking to discredit the quality and reliability of the evidence (Pope, 1995). The drive to discredit

someone else's research could come from a strong concern or judgement call, but it could also stem from the critic's own biases[11](Ortlipp, 2008).

Statistics Reporting and Research Bias in Practice: The Harvard Sugar Investigations

To help argue in favor of a framework for research ethics that is inclusive of methodological decision-making, I established two underlying factors influencing replicability of scientific studies (particularly in the social sciences): manipulation of statistics reporting and methodological reporting bias. Until this point, I discussed both the reporting and bias from a conceptual and didactic standpoint, with the intent to highlight and explore the concepts and their characteristics. It is important, however, to understand these concepts through a concrete example. Thus, I will explore one historical case in which both manipulation of statistical reporting of data and researcher bias influenced the outcome of important work: namely, the Harvard Sugar Studies.

Harvard Sugar Studies

For decades, the Corn Refiners Association (CRA) in the US funded Harvard scientists to study the chronic effects of sugar consumption by adults, children, and animals (Kearns, Schmidt & Glantz, 2016). Findings from the investigations downplayed sugar's role in negative health outcomes such as obesity, diabetes, and other co-morbidities. Instead, evidence pointed to other dietary and lifestyle variables,

[11] One practical recommendation to add credibility to qualitative work, contends Ortlipp (2008), is to keep a detailed field journal and document actions during all stages of an investigation.

such as dietary fat and increased sedentary activity, as actual underlying mechanisms associated with overweight and obesity. Thus, from the 1970's until the early 2000's these Harvard-led investigations dictated the direction of food-based policy (e.g. the carbohydrate-driven food pyramid) and dietary fads (e.g. the low fat diets) which were marketed to the public as cornerstones to living a healthy life (Brinton, Eisenberg & Breslow, 1990).

Conversely, when future investigators, free of CRA funding, began conducting replication trials of the original investigations, they found their results not only clashed with original reports but also were consistently demonstrating the findings from older investigations were untrue. Rather than having only a minimal role in negative health outcomes, these newer studies found high sugar intake linked to increased morbidity and diet-related diseases such as diabetes (Kearns, Schmidt & Glantz, 2016). Today, not only has the influence from the Harvard sugar studies waned, but newer diet trends and food-based policies are in place that reflect the shift in findings generated by the newer trials (e.g. low-carbohydrate diets, CDC food plate).

Perhaps the most blatant answer to the question of *why* the initial Harvard Sugar Studies and subsequent replication trials differed and contradicted each other is that one of the two groups was biased, either intentionally or unintentionally. Critics of the investigations, or of the CRA, are quick to highlight: because the Harvard investigators were funded directly by the CRA and the Sugar Industry, they were expected to produce findings that were in line with the industry's agenda. However, is it fair to accuse one group of investigators of funding bias when the principal

investigators of the replication studies may have been intentionally or unintentionally biased *against* sugar, themselves?

Though funding likely plays a large role in shaping the outcomes of an investigation (Lexchin, 2012) there is no other guarantee, or check, that can definitively say if either group is biased apart from relying on the current subjective verdict-driven judgement. However, under the scenario in which blame is allocated with underlying and undiscussed subjectivity, it is unlikely a fair and equitable resolution will ever be reached. The investigators on the Harvard research team are very unlikely to agree, let alone admit, their work was biased despite being funded by industry monies. The same can be inferred for the replication scientists— they will not view their work as biased, but as credible science that produced quality results contradicting the original investigations' findings.

Objective Assessment Tools for Methodological Reporting and Bias

At a bare minimum, scientists need methodological tools to assist in mitigating subjectivity in bias detection, to prevent more cases similar to the Harvard Sugar Studies. As it presently stands, bias is something that is subjective and, consequently, very difficult to capture due to clashing opinions of scholars and a conceptual gap in understanding of how bias affects research. If subjectivity were removed (or minimized), the (a) abilities to confidently identify bias, and (b) the procedures regarding how to handle biased work could significantly improve.

In research, scholars deal with two components in the reporting of their studies: (a) numbers, and (b) words. I contend that both components serve as potential domains

in which intentional and unintentional biases can manifest themselves. Henceforth, I will refer to these, as "numeric and language bias"

Numbers and words are potential avenues for bias because one can manipulate numeric data or spin verbal arguments with flowery language, deliberately, to make the results of testing a hypothesis more attractive. Misrepresentations of numbers and words also can occur due to unintended bias—poor statistics education can lead to unintended error during the application of methods or presentation of findings. With word bias, we may unknowingly be using language patterns and structures to explain something incorrectly, or not be providing sufficient information to articulate an accurate representation of an idea.

Today, technology can assist with strategies that were previously impossible, for mitigating both numeric and word bias. Across multiple fields outside applied sciences, computer programmers have created various tools and algorithms intended to examine numbers and language from a more objective standpoint. In other words, rather than have researchers rely on subjective frameworks to detect bias, they now can rely on computers to assess quality of findings and to mitigate subjectivity.

If these tools allow for less subjectivity in bias detection, then perhaps applying them in Public Health research could help address research ethics issues. Therefore, below I briefly discuss: (1) assessment tools for detecting potential numeric bias, and (2) assessment tools for detecting potential language bias. The intent of this discussion is to highlight the underlying mechanics of these new technologies and to

argue for the adoption of these tools as legitimate strategies to aid in bias detection and mitigation.

Assessment Tools for Numeric Bias

Numeric bias can be categorized as any type of inaccurate representation of numeric data. Much of what is driving the replicability crisis is poor representations of data. Determining if data and their analyses are inaccurate, however, is a daunting task for any reviewer and involves full exchange of datasets— something many scientists would not, willingly, agree to (Corlett, 2005). Assessment tools for numeric bias not only make the task of finding errors in published work easier, they also serve as a validity check for anyone seeking to guarantee their work is free from error. Specifically, I will elaborate on two useful tools: (1) statcheck (*sic)* and (2) the Grading of Recommendations Assessment Development and Evaluation (GRADE) scale.

Statcheck [*sic*]. statcheck[*sic*] is a free, downloadable program for the open-source statistical programming software R, designed to re-calculate and, if necessary, correct, *p*-values in published manuscripts (Nuijten, 2017). Today, statcheck's users rely on the program as a valid tool to check the accuracy of their reported findings. Though useful at mitigating unintended error by helping users report accurate *p*-values, statcheck's programmers (Epskamp & Nuijten, 2015) contend their program has far greater capabilities, arguing:

> Conclusions in experimental psychology often are the result of null hypothesis significance testing. Unfortunately, there is evidence that roughly *half* of all published empirical psychology articles contain at least one inconsistent *p*-

> value, and around one in eight articles contain a grossly inconsistent *p*-value that makes a non-significant result seem significant, or vice versa. Often these reporting errors are in line with the researchers' expectations, which *means these errors introduce systematic bias*…[therefore] statcheck can be used to evaluate the prevalence of reporting errors (Nuijten, Hartgerink, Van Assen, Epskamp, & Wicherts, 2016, emphasis added) (Retrieved from: https://mbnuijten.com/statcheck/.

To test the validity of statcheck as a tool for detecting number bias, Nuijten & Colleagues (2016) used the program to replicate an exhaustive study by Wicherts et al. (2011), which manually analyzed the accuracy of reported significance tests from eight major psychology journals ranging from 1985-2013. Findings from Nuijten & Colleagues' (2016) replication investigation found statcheck was able to successfully detect *most* of the same errors found in the manual version of the study. More importantly, "…statcheck found an inconsistency rate that was 4.7 percentage points higher than the one in the manual search" (p. 122).

One of the most salient criticisms raised against statcheck is, in its current version, that the program can only analyze results from manuscripts formatted according to the APA. That is, if the study is published in any different reporting style (e.g. American Medical Association, Chicago, among others), the algorithm will overlook that reported *p*-value. The program's reliance on APA formatting is, by admission of the authors and programmers, perhaps the main reason why statcheck was unsuccessful at extracting *all* significance tests in their replication of Whichert's et al.

(2011) investigation. However, those authors still managed to a large number of inconsistencies within Whichert's investigation.

Regardless of this limitation, statcheck should be considered a useful tool in Public Health research. As with Psychology, Public Health and other Health-related academic fields often over-rely on significance tests to report results (Barry, et al., 2016). Many flagship journals, namely the *American Journal of Public Health* and *Health Education and Behavior*, mandate that all manuscripts be presented in accordance with the recommendations from the APA. Thus, statcheck can readily be applied to journals currently utilizing APA format. A second practical remedy to statcheck's limitation is to, theoretically, re-write the text to fit APA style by hand. Though potentially arduous, this approach can be utilized for other journals that do not rely on APA format.

GRADE[12] scale. The Cochrane[13] Institute's GRADE Scale, a mandatory component of all Cochrane Systematic Literature Reviews, is a series of parameters intended to objectively 'grade' the quality of a body of evidence. Once calculated by hand, today Cochrane's online software interface (GRADEProGDT[14]) allows users to input numeric information on studies testing the same hypothesis (e.g. 'medication X is

---

[12] The section is meant to provide a brief overview of GRADE, the GRADE Scale, and online program GRADEpro. A full treatment of GRADE can be found in Chapter 3.

[13] Cochrane is a non-profit, non-governmental organization dedicated to evaluating medical research to facilitate strong, evidence-based decisions about health interventions http://us.cochrane.org/.

[14] A Link to GRADEProGDT can be found at: https://gradepro.org/

more effective than a placebo'). Extracting the numeric data from each study produces what is called a summary of findings table (SOF). The goal of the SOF table is to determine how precise the presentations of the data are, by assigning a 'grade' to the evidence's quality: (1) High Quality, (2) Moderate Quality, (3) Low Quality, and (4) Very Low Quality.

Numerous criteria can improve or harm a study's final 'grade'. In an investigation titled: *What is 'Quality' of Evidence and Why is it Important to Clinicians?* Guyatt & Associates (2008) elaborate on what, exactly, elevates or downgrades evidence quality within the GRADE scale framework, contending:

> [1]Randomised (*sic)* trials begin as high quality evidence and observational studies as low quality evidence. [2] Quality may be downgraded as a result of limitations in study design or implementation, imprecision of estimates (wide confidence intervals), variability in results, indirectness of evidence, or publication bias. [3] Quality may be upgraded because of a very large magnitude of effect [effect size], a dose-response gradient, and if all plausible biases [were controlled for]. [4] Critical outcomes determine the overall quality of evidence (p.996).

Since the establishment of the GRADE scale in 2002, researchers have employed it to test the validity of findings in studies across several academic disciplines. Many researchers (e.g. Guyatt & Oxman, 2016; Atkins, Best, Briss & Ecceles, 2004; Zhao, Liang, Fang & Liu, 2017) find the GRADE scale is better able to capture discrepancies in the presentation of findings than traditional reviews alone.

Puhan and colleagues (2014), for example, used GRADEpro software to analyze the cohesiveness of treatment effect estimates from meta-analyses in medicine on the subject of hospital networks. The authors concluded many of the analyzed studies were either inconsistent or missing critical values in their presentation of findings. More important was Puhan and colleagues using their findings to provide an objective assessment and practical recommendations on data reporting, contending:

> following the [five] steps[15] [in the GRADE framework used to assess evidence quality] highlights the necessity for authors of NMA to present direct, indirect, and NMA estimates as well as quality ratings for all direct comparisons. If authors do not present these estimates, skepticism regarding any inferences from the NMA is warranted. (p.5)

The GRADE scale, however, is not immune to detractors who argue GRADE, in and of itself, is flawed. Kavanagh (2009) states, simply, the requirements of the GRADE scale are too complicated to be internally and externally consistent. Kavanaugh continues, there are several challenges that inhibit GRADE's ability to detect inconsistencies (e.g. language barriers, international laws, international research standards). Therefore, contends Meader et al. (2014), there should be a check list to aid

---

[15] The five steps in the GRADE scale used to evaluate quality of the evidence are: (1) Study design (e.g. randomized control, observational, among others) (2) Risk of Bias, (3) Indirectness (Did the study measure what it intended to measure?), (4) Imprecision (Were hypotheses tested in the same manner?) and (5) Inconsistency (Were findings consistent in all studies?). Depending on the answers, researchers cumulatively upgrade or downgrade the overall quality of the evidence as being either: (1) high, (2) moderate, (3) low, or (4) very low.

consistency and reproducibility of GRADE assessments. More importantly, more replication investigations of previous GRADE systematic reviews are needed to ascertain levels of replicability.

Because the nature of Public Health and other health related fields' research involves human subjects, clinical trials, and large-scale interventions, the GRADE approach should be considered a viable component of systematic reviews and meta-analyses. Through GRADE's record of success, researchers can effectively assess the quality of a collection of numeric findings. The ability to assess the quality of findings holds important implications for the future direction of Public Health research—namely, the ability to ascertain quality through more objective frameworks. Despite its inherent limitations, employing the GRADE scale could eventually contribute to stronger reporting practices and more transparency in methodological decision-making in studies conducted in the future

In sum, while detection of numeric bias is inherently subjective, advances in technology are beginning to provide tools to capture this type of bias, more objectively than before. These two tools: (1) statcheck, and (2) the GRADE scale, rely on technology to extract, analyze, and recalculate evidence to assist with objectively assessing the validity of the data. I argue they should at least be explored further in Public Health research as a means of mitigating subjectivity in numeric bias detection, and improving the overall quality of research reporting.

Assessment Tools for Language Bias

Word bias is defined, here, as intentional or unintentional use of vocabulary and word patterns to sway interpretation of an outcome. Unfortunately, detecting language bias may be even more difficult than number bias, as accusing scholars of using false, or misleading words can be viewed as inflammatory behavior. Scholars, however, have tried. For instance, Egger and associates (1997) found when randomized control trials in Germany were translated into English, the language used in the English version overly-emphasized significant findings and often ignored important non-significant findings, or adverse effects. Egger contends this unethical behavior occurred, most likely, to secure publication in top journals within the United States. Though Egger had evidence of potential bias, little else could be done to stop the practice from occurring due to, at the time, limited understanding of *how* to handle biased language in a non-subjective manner.

Unlike numeric data, which can be checked, re-analyzed, and corrected, it is much harder to 'fix' language patterns, if they are deemed biased (Mescasens, Danescu-Niculescu-Mizil & Jrafsky, 2010). Technology has been advancing to a point, however, where programmable computer algorithms can 'map' and 'dissect' language to more objectively determine if certain latent word patterns reflect embedded bias (Hu, Boyd-Graber, Satinoff & Smith, 2014). In the subsequent section, I will discuss a technique in Computer Sciences, named Topic Modeling (TM), and elucidate why TM holds important utility for language or word bias detection in Public Health.

Topic Modeling. Topic modeling refers to a series of algorithms that use matrix algebra to map latent language patterns from a large collection of text documents[16] (known as a *corpus*) (Blei et al., 2001). Primarily used in Computer Science and Business Marketing, TM has been widely regarded as the premiere methodology for consolidating, mapping, and assigning structural meaning to an otherwise insurmountable amount of online data (Wallach, 2006).

In Business Marketing, major retailers such as Amazon, seek to understand what/why customers prefer (or not) a popular product. To answer their query, the marketing team would likely turn to product reviews for the necessary information to reach an informed conclusion. Sorting through tens of thousands of reviews, however, is both time consuming, incredibly difficult for most humans, and most importantly for our purposes, subjective. The TM algorithm, however, can process these reviews in a matter of minutes, to create a model with a pre-specified number of 'topics' and a collection of words most commonly associated with each topic (Suominen & Toivanen, 2016). In other words, rather than sort a large mass of text manually, the algorithm has learned a general 'snapshot' of the most important pieces of information on what the full corpus represents. The marketing team can, now, reach an informed conclusion by interpreting a mathematical 'reduction' in the corpus rather than sifting through every product review.

---

[16] This section is centered on providing a conceptual overview of topic modeling. A full treatment of the mechanics in TM can be found in Chapter 4 of this dissertation.

The process of topic modeling is surprisingly straightforward. First, one needs a large collection of text data— which can be anything from textbooks, social media posts, personal essays, or other formatted text. The text data are saved as one file, and fed through the statistical software program R. Once in the program, the text goes through a series of changes to prepare for analysis— e.g. punctuation, as well as unimportant words such as articles and suffixes are removed to make the corpus more concise. Once the data are cleaned, they are analyzed through an iterative Bayesian process that compares each word *x* to every other word *y* in the entire corpus.

Words with high degrees of association (e.g. word *a* has a high probability of appearing with word *b*) are clumped together to form a latent theme, or topic. The clusters of words in each topic should be clear and interpretable and representative of a hidden theme within the broader corpus. For example, if we generated a topic model from books on international cuisine, we would likely identify one topic as 'American' or 'Italian' based on word clusters in each latent topic.

To demonstrate TM and argue for the application of this methodology in social sciences, Valdez, Pickett & Goodson (in-press) generated a topic model of the transcriptions for the 2016 Presidential Debates. The purpose for using the Presidential Debate transcriptions as a corpus was to answer two specific research questions: (1) Would emergent topics align with policy related concerns ahead of the presidential election, and (2) How differently do politicians react to the same question? Valdez et al. found that (1) the emergent policy-related topics aligned with the most highly searched Google items one week before the start of debate one, and (2) each

candidate's respective topic model was composed of similar topics (i.e. policy related matters) but differed significantly in the words used to compose that topic.[17]

As with any methodology, there are a series of limitations associated with TM. First, topic modeling is a divisive tool, causing debate over *which* algorithm works best. Currently, the two most widely used and recognized algorithms are the Latent Dirichlet Allocation, and Latent Semantic Analysis. While both algorithms perform the same task, they arrive at their conclusions via different mathematical calculations— LDA relies on Bayesian inferencing, and LSA on eigenvalues. Despite providing almost identical outputs (e.g. Anaya, 2011), LDA is widely preferred over LSA because the algorithm is driven by probability distributions and not GLM type analyses (Bergamaschi & Po, 2011). Unfortunately, conducting the more popular LDA requires familiarity with statistical software program "R", which is sometimes viewed as a tedious resource by those unfamiliar with the language[18].

A second limitation of TM is that, even in cases with the most clear and interpretable of outputs, the researcher must ultimately define and assign meaning to each topic— a process similar to what is done in exploratory factor analysis (EFA)

---

[17] For example, both Candidates Clinton and Trump had a "Second Amendment" topic in their respective topic models. Clinton's Second Amendment topic included the following words ("gun", "second amendment", "loophole", "close") and Trump's included words such as ("Second Amendment", "Supreme Court", "appoint", "protect", "guns"). Thus, the computer successfully mapped out the same topic for each candidate, but identified different words – and, therefore, different perspectives -- addressing each topic.

[18] LSA has a point and click interface on SAS.

(Gorsuch, 1988). In EFA, the researcher merely has 'recommended cut off values' to determine how many latent factors emerge among a collection of measures. In TM, we have a collection of seemingly similar words grouped together into latent topics. In both cases, the researcher must determine commonalities among the grouped measures/words and assign qualitative meaning to them.

Despite these limitations, TM should be viewed as a legitimate tool in Public Health. Specifically, TM can be used to examine how certain variables, such as time or funding source, could influence and potentially bias the wording in published manuscripts. For example, how has language regarding HIV/AIDS in published studies changed from the start of the AIDS crisis in the 1980's to today? Or, do manuscripts testing the same hypothesis result in different topics when they are funded by industry versus federal government grants (recall the Harvard Sugar Studies)?

In the HIV/AIDS example, the language embedded in published investigations, most likely has become softer, shifting away from pandemic language toward treatment, chronic disease management, and self-care. For the sugar example, the answer would likely be that studied funded by the sugar industry and by federal government sources might arrive at different conclusions because funding source is frequently cited as a bias-inducing factor (Barden, Derry, McQuay & Moore, 2006).

By generating topic models for each group (e.g. topic models on sugar research funded by industry versus federal funding) researchers have a unique opportunity to compare these mathematically supported models to determine if the language suggests potential bias(es). For example, if the topic model for the studies funded by federal

monies used extremely critical language against sugar (e.g. *"sugar", "bad", "adverse", "sick", "metabolic", "disease"*) but the industry model referred to sugar favorably (e.g. *"sugar", "fine", "intake", "juice", "nutrition"*), one could conclude the industry-funded studies are biased because it is generally accepted that high sugar intake is unhealthy

Concluding Remarks on Research Ethics

While in the academic world trained scholars can judge the quality and merit of scholarly work, the lay public cannot. Instead, the lay public relies on the practical, evidence-based recommendations from authoritative sources to guide their own lifestyle choices. But when bias plagues academic work, and current attempts to address bias in academia lead to finger-pointing and contentious backlashes, the lay public may not be receiving the best or most accurate advice when research guides legislation and policy development.[19]

Even for those who contend methodological decision-making is *not* a major ethics issue, one cannot deny providing misleading information to the general public is a detriment to the public good. Therefore, at the bare minimum, at least acknowledging something in the system requires reform is key to progress. Until policy is guided by

[19] On February 2nd, 2018, Raw Story reported corn syrup lobbyists were actively working with the current presidential administration to set USDA dietary guidelines citing research that downplayed negative health outcomes of increased high-fructose corn syrup (HFCS) consumption.

work that has minimum bias[20], scholars should continue to pursue reform, or at the very least, uphold integrity when producing science.

The potential for harm to the general public via misinformation or public policy, and the morality of providing sound science, are precisely the reasons why methodological decision-making should be viewed as an extension of research ethics. Though the rights of animals and participants are inherently important, the proliferation of dubious science could pose equal, if not greater consequences both to the scientific enterprise and to the lay public. Therefore, when discussing, or debating, the practice of research ethics, methodological decision-making should be viewed as an issue that is, at the very least, *as* important as human or animal rights.

In this investigation, I argued there are two factors contributing to methodological decision-making as an ethical issue: (1) manipulation of reporting, and (2) researcher bias. Both factors can be either unintentional or deliberate— however, neither is worse or more harmful than the other. By understanding *how* these factors drive issues within the research system, we now have a conceptual understanding regarding the extent to which research is being harmed by the continued presence of underlying research bias and by the practice of poor methodological reporting habits.

Fortunately, to mitigate bias, technology has furnished increasingly reliable tools for assessing bias more objectively. Within two particular domains of bias – numeric bias and word bias – I described three potentially useful tools to assist in bias

---

[20] Note: because it is a socially constructed phenomenon, science will never be completely bias-free.

detection: (1) statcheck (2) GRADE scale, and (3) topic modeling. Although each method has its own inherent limitations, their utility in Public Health and other health related fields is apparent. If these tools can potentially identify and reduce bias in research, they should, at the very least, be tested to examine the extent to which (1) the tools, themselves, are valid and (2) whether they can, indeed, capture and help mitigate bias in Public Health research.

Although none of these methodological tools are new, they are rather novel to Public Health and other health related fields. Therefore, I would strongly encourage that, moving forward, scholars use and apply these tools to evaluate their own research— even if they contend their work is free of bias. By engaging with these tools, and reflecting on their own research practices and habits, researchers can work to become more informed about bias and, hopefully, contribute to viewing and addressing methodological decision-making as a vital ethical issue for the scientific enterprise.

CHAPTER III

NUMERIC BIAS AND GRADEPRO

In 2017, after allegations of false peer-review and questionable reporting practices, publishing group Springers retracted 107 articles on innovative cancer pharmacotherapies from its top-tier Oncology research journals. The retraction stunned many in the scientific community (Stigbrand, 2017). More importantly, though, were the subsequent discussions regarding the merit of other investigations testing new pharmacotherapies intended to contribute to the benefit of the public's health.

With these discussions came the realization that Springer's retraction is not a one-time phenomenon. It is, nonetheless, the *largest* instance of manuscript retraction in the history of academia. Springer's example is the culmination of a growing concern in the research communities in the US, beginning in the 1970's: increased instances of manuscript retractions. Over the last decade, for example, manuscript retraction rates have increased ten-fold (Fang, 2011).

There are at least two ways to interpret this increase in retractions of published scientific reporting: either (1) the self-correcting nature of science is improving, or (2) scholars are abusing the system and putting forth poor quality, biased work (Cokol, Ozbay & Rodriguez-Esteban, 2008). To better understand the retraction phenomenon, Fang, Steen & Cassadevall (2012) elected to review the 2,047 articles retracted by PubMed from 1973-2011. Those authors' aim was to determine whether manuscripts were retracted due to underlying bias, or honest error. Results from their analysis

indicated that, in 33.6% of cases, retraction was due to unintentional mistakes in data reporting. Conversely, the remaining 67.4% of cases were retracted due to fraud, suspected fraud, duplicate publication, or plagiarism. The findings from Fang and colleagues fueled the concern over the extent to which bias is present in published work, but remains undetected (Van Noorden, 2011). However, regardless of the reason a manuscript was retracted, Fang and colleagues ultimately concluded, intentional or not, the primary culprit behind retracted studies was biased data.

The purpose of this investigation, therefore, is to raise awareness regarding the potential for number/numerical bias (cf. definition below) in large-scale health/drug studies. Specifically, in this chapter, I will argue for the importance of detecting numeric bias in academic/scientific work, by addressing: (1) how numeric bias can manifest itself in research, (2) the effect of numeric bias on public policy, (3) innovative approaches/tools to identify numeric bias and (4) a heuristic example using GRADEpro to assess potential bias in large-scale clinical trials of PrEP (Pre-Exposure Prophylaxis, an initiative to promote the use of HIV treatment drugs for preventing HIV infection, and one of the latest pharmacotherapies available in the United States).

Numeric Bias in Research

Numeric bias was defined in the previous chapter as any intentionally or unintentionally inaccurate representations of numeric data in scientific research (Munafò et al., 2017). When referring to number bias, there are generally two categories: (1) unintentional numeric bias, and (2) intentional numeric bias. Unintentional numeric bias is best described as honest errors made by a researcher at

any stage of an investigation (Cook, 2014). A researcher may, for example, accidentally report the wrong *p*-value, or run an analysis that is inappropriate for the data at hand (e.g. conducting a t-test when there are more than two groups). Conversely, intentional bias refers to the direct manipulation of data to assist in falsely supporting a hypothesis. Common examples include: (1) adding, deleting, and carefully selecting cases to analyze, (2) purposefully reporting false findings, and (3) inflating sample sizes.

Regardless of intention, bias harms research in various ways, but especially through a decline in both the quality of scientific investigations and the general validity of findings. Ioannidis (2005) contends, for example, numeric bias has become so commonplace in the academy that, today, most findings in published research may be false. Hubbard (2014) adds that all academic disciplines are prone to numeric bias. He further cautions that so long as biases go unchecked, the quality of research, as a whole, will reach increasingly new lows.

Numeric bias is difficult to capture, for various reasons. Chief among them, is the sheer volume of scientific publications in professional journals— estimated at 27,000 articles published every week— which allows for increased levels of questionable, or biased work to pass through the peer review system (Van Noorden, 2011). Due to such proliferation of published studies, among other factors, the academic community cannot escape the reality there could be a potential correlation between increased volume of accepted manuscripts and increased instances of numeric bias.

One manifestation of the link between publication volume and number bias is the amount of journal article retractions. While once a rare event, today the process of retracting an article from an academic journal is much more common (Finelli, 2013; Hesselmann, Gar, Schmidt & Reinhard, 2016). As mentioned above, within the last decade, retraction rates have increased ten-fold. In the 1970's, for example, retraction rates from journals across academic disciplines nearly *quadrupled* despite increased editorial oversight and journal-wide reporting recommendations intending to improve study quality (Finelli, 2013). Granted: retractions are indicative of the self-correcting nature of science (Marckman, 2010), but their massive increase is fueling concerns regarding *why* poor research still permeates a system intended to uphold scientific integrity (Wiles, 2014).

Along with fast proliferation of manuscripts, a second reason number bias persists in research, observe Ortega & Navarette (2017), is the hegemony of significance testing— which is the dominance of null-hypothesis significance testing as the primary method to relay findings (Nuzzo, 2014). Barry et al. (2014) for example, reviewed 1,245 articles from flagship research journals to assess reporting practices of each study published between 2000 and 2012. They (Barry and colleagues) found that in 100% of the cases, significance tests were the primary statistic used to communicate findings, chosen more frequently than other more informative statistics such as effect sizes.

The continued reliance on null-hypothesis significance testing (over other vital statistics such as confidence intervals and effect sizes) has led to some critical

implications in published research, namely: (1) bias against non-significant findings in academic journals (Easterbrook, Gopalan, Berlin, & Matthews, 1991; Csada, James, & Espie, 1996), (2) decreased use of other statistics that measure study quality (Calin-Jageman, 2017), (3) decreased replicability of study findings (Loken & Gelman, 2017), and (4) public mistrust of research due to lack of reproducible findings (Scharff et al., 2010). Therefore, if significance testing remains the primary reporting tool, there will always be a risk of bias in research (Stern & Smith (2001).

Though poor reporting practices and fast proliferation of manuscripts challenge the production of quality science[21], researchers' continued reliance on outdated methods such as significance testing is emblematic of a larger issue within all academic disciplines— namely, the extent to which number bias is potentially embedded in literature, yet scantly detected or addressed, until long after a questionable study has been published (Young, 2009). Consequently, much of the 'suspect' work persists as credible evidence, to be read, cited, sourced, and used as support when forming evidence-driven policy or treatment-related decisions (Gino & Bazemann, 2009).

The Effect of Number Bias on Public Policy

To propose new policies supporting the public's health, policy makers must look to scientific literature for a platform on which to anchor their recommendations ("From

[21] "Quality research...pertains to the match between the methods and questions, selection of subjects, measurement of outcomes, and protection against systematic bias, nonsystematic bias, and inferential error" (National Center for the Dissemination of Disability Research, 2018, retrieved from: http://ktdrr.org/ktlibrary/articles_pubs/ncddrwork/focus/focus9/Focus9.pdf)

the Science Policy Blog.," 2009). Thus, findings from scientific investigations, published in the form of peer-reviewed research reports, greatly influence national Public Health policy and, ultimately, the health behaviors/outcomes of entire populations. Put bluntly, research guides policy, and policy guides the public.

Most policy makers are aware of this fact. What they may not be as keenly attuned to, however, is that reliance on published research ultimately binds Public Health policies to all other features of the studies generating those findings, such as: the methodological and analytical decisions made by the researchers, researchers' personal or professional biases, the idiosyncratic weaknesses of each study, and the potential inaccuracy or purposeful manipulation of a study's findings. (Tunis, Stryer, & Clancy, 2003).

Policy makers and the lay public, however, are not trained to evaluate the mechanisms underlying scientific research and, thus, are ill-equipped to distill quality science from problematic science, which is often infused with confusing language and scientific jargon (Popay & Williams, 1996). Therefore, non-scientists who turn to scientific literature to make informed decisions could be scaffolding support with faulty material.

There are several instances in the history of Public Health in which research of poor quality was used as credible evidence by the lay public and policy makers. One case, for instance, occurred in 2008, when decades of biased research were used as evidence to approve a mandate that hospitals and doctors spend trillions of dollars on expensive high-technology databases (known as the HITECH law) to mitigate rising

patient mortality (Soumerai & Koppel, 2017). However, almost all evidence used to support the HITECH law relied on poor research designs (t-tests comparing death rates between high-tech versus low-tech hospitals, for instance) that failed to control for confounding variables such as patient and regional neighborhood wealth and wellness (i.e. healthy lifestyle habits, such as diet and exercise), among others.

The lay public and journalists also share blame for unintentionally sensationalizing misinformation derived from poor research. For example: Several studies reported between 2010-2016 found a link between newer, advanced-life-support ambulances, and *increased* risk of mortality when compared to basic ambulances (Sanghavi, Jena, Newhouse & Zaslavsky, 2016). Headlines were quick to broadcast that link, which prompted activists to lobby Congress to fund research that further explored the link between ambulance quality and mortality (Kane, 2014). The studies, however, relied only on anecdotal evidence— and failed to disclose that newer ambulances were dispatched only in situations where the patient was five times more likely to die of a life-threatening condition (Soumerai & Koppel, 2017).

Kairney & Oliver (2016) contend that to prevent the spread of misinformation, policy makers *should* work in tandem with scientists to forward policy-agendas. However, those authors further state, there is limited evidence on *how* to streamline communication among scientists, policy makers, and the lay public. In the Institute of Education's 2016 report on the science of using science, Langer, Tripney and Gough argue that the barrier to communication between policy makers and scientists stems from professional differences and foci. Specifically, scientists spend weeks, months,

and even years drawing objective conclusions based on collected data and tested hypotheses to produce knowledge. Legislators, journalists, and the lay public on the other hand, are users of that knowledge and may not engage with the content as critically or as objectively. Chan and Altman (2005) add that the relationship between scientists and policy makers is further complicated when sensationalized headlines or news titles clash with lackluster titles from scientific reports.

Medical ethicist Altman (2002) argues that, due to the spread of poor-quality information and general inability to find consensus on *what* represents quality work, all research should undergo a systematic process to parse out poor-quality science. Creating and implementing a systematic approach to detect bias, however, should not fall to the lay public and policy makers, but to scientists, themselves (Altman, 2002).

Innovative Approaches/Tools to Capture Number/Numerical Bias

Increased retraction rates, the impact of poor-quality research on public policy and the public's wellbeing, alongside heightened awareness of number bias have sparked renewed interest in the study of bias, among scholars. More importantly, scholars find themselves increasingly determined to create means/methods to: (1) better detect bias, and (2) promote transparency within research, with the ultimate purpose of bolstering the credibility of scientific investigations.

Marcus & Oransky (2014), for instance, claim one such tool prompting renewed interest in promoting quality science is their own initiative, Retraction

Watch,[22] which serves as an academic watchdog to notify when an editorial board has retracted a study from publication, and for what reason(s). Though certainly not without its detractors (e.g. Teixeira da Silva, 2016), platforms such as Retraction Watch are credited, primarily, with promoting transparency in science by providing a better glimpse of (1) how the scientific process operates, and (2) whether science is operating according to strong ethical standards.

While Retraction Watch and similar venues have renewed and improved bias vigilance in academic research, these can only *inform* readers about instances of retraction. In other words, Retraction Watch cannot *capture* bias, but instead relies on reports from journal editors when they post a notice of retraction. Thus, Retraction Watch is inherently limited regarding what the initiative can contribute. Today, however, scholars can utilize various tools to scan printed text to determine if the presentation of findings contains embedded bias.

GRADEpro and "statcheck" (*sic*) are two such tools that rely on computer-based technology to reach evidence-driven decisions about the quality of the numeric evidence in scientific investigations (Kavanagh, 2009; Meader et al., 2014; Puhan et al., 2014; Nuijten, Assen, Hartgerink, Epskamp, & Wicherts, 2017). GRADEpro, for example, relies on a set of pre-specified criteria to up-grade or down-grade the quality of evidence supporting tested hypotheses in studies. The specified criteria, intended to find flaws within study procedures are: (1) Risk of Bias, (2) Imprecision, (3)

---

[22] The website can be found at:www.retractionwatch.com

Inconsistency, (4) Indirectness, (5) and Publication Bias. Depending on the overall evaluation score on each of those domains, the tested hypothesis receives one of the four following grades: (1) High, (2) Moderate, (3) Low, and (4) Very Low.

"statcheck" (*sic*) uses an algorithm to explore numeric data (i.e. significance tests, the corresponding F-statistic, and any confidence intervals) to determine if the reported *p*-value on a given significance test is accurate or inaccurate. If the *p*-value is inaccurate, the algorithm will recalculate the reported finding and provide a corrected *p*-value that reflects the reported F-statistic and sample size. The final output generated from statcheck provides an overall assessment on how trustworthy the results are, based on the ratio of accurate to non-accurate *p*-values.

An added benefit of platforms like GRADEpro and statcheck is their validation via research and subsequent endorsements by the scientific community. Harbour & Miller (2001), for example, propose that GRADEpro's consistent, systematic approach to evaluating literature make it an attractive tool to enhance systematic literature reviews and meta-analyses— especially as GRADEpro becomes more widely used and adopted. Baker (2016) adds, while programs like statcheck are new, "in the long run [they] could keep scientists honest [especially] if researchers made raw data available" (p.151).

With the growing call for transparency in science and increased availability of tools intended to address bias such as GRADEpro and statcheck, scholars are better equipped to study and mitigate bias than ever before. And, though GRADEpro has been applied has been widely applied to systematic literature reviews to evaluate

evidence quality, the tool has yet to be applied to a Public Health audience. Therefore, testing the tool with a timely, data-driven heuristic example could elucidate whether GRADEpro is an adequate tool to add to the bias-detection arsenal.

A Heuristic Example: Using GRADEpro to Assess Potential Numeric Bias in Clinical Trials

Background

In December 2011, the journal *Science* named "HIV Treatment as Prevention" its *Scientific Breakthrough of the Year* (Alberts, 2011). This "HIV Treatment as Prevention" initiative — known as pre-exposure prophylaxis, or PrEP— comprises a fixed-dose combination of two anti-retroviral drugs: tenofovir disoproxil fumarate and emtricitabine. A single pill combining both drugs— popularly known by its brand name Truvada— received FDA approval in 2004 for treating HIV, and in 2012 as a prophylactic measure to prevent infection (CDC, 2017, retrieved from: https://www.cdc.gov/hiv/basics/prep.html). As PrEP, physicians prescribe Truvada to persons who test negative for HIV but are at high risk of contracting the virus (Galea et al., 2011).

Despite the novelty of PrEP, treatment as prevention (i.e., the main application of PrEP) is not a new phenomenon. There are numerous examples of treatment as prevention in the medical and Public Health literatures; including: (1) daily aspirin regimen for heart health (Juul-Moller et al., 1992), (2) psychiatric medication for mental health in patients with a documented family history of mental illness (Jellinek,

2003), and (3) prescription-strength multi-vitamins for preventable diseases, such as osteoporosis (Ooms et al., 1995).

As with other medications and therapies, PrEP's promise to improve quality of life easily captures the attention of HIV activists and policy makers who aim to make the medications more accessible through public policy recommendations and other government-related initiatives (Bongaarts & Over, 2010; Fauci, 2011; Cockcroft, Masisi, Thabane, & Andersson, 2014). For example, shortly after PrEP received approval from the Food and Drug Administration (FDA), the National Alliance of State and Territorial AIDS Directors (NASTAD) released a policy statement (2012) on PrEP arguing, in part:

> The opportunities afforded by PrEP are unprecedented in the Public Health response to the epidemic. The daily utilization of Truvada as a mechanism to prevent HIV acquisition would allow for an individually-controlled, moderately effective prevention tool that [should] be used alongside other proven prevention methods, with or without the knowledge and cooperation of a sexual partner. In the scope of prevention science, it may be the closest we have come to a vaccine. (Retrieved from: https://www.nastad.org/sites/default/files/resources/docs/PrEP-Policy-Statement-FINAL-6.25.12.pdf)

In 2015, support for PrEP reached new levels when for the first time it was addressed in the US President's Emergency Plan for AIDS Relief (PEPFAR). Particularly, a PrEP regimen was billed as a legitimate prevention drug that needed

immediate prioritization among HIV high-risk groups, including: (1) international young women geographically in the 10 PEPFAR DREAMS[23] countries with high HIV prevalence, (2) HIV serodiscordant couples, (3) female sex workers, (4) men who have sex with men, and (4) people who inject drugs. PEPFAR's stance on PrEP concludes:

> …with *strong evidence for the efficacy and effectiveness* of daily oral PrEP across multiple studies, *it is a* Public Health *priority* for PEPFAR to make PrEP available in high HIV prevalence settings in a strategic fashion to people at substantial risk, including adolescent girls and young women, HIV serodiscordant couples, female sex workers, men who have sex with men, and injection drug users, based on their risk. (p.2, emphasis added) retrieved from: https://www.pepfar.gov/documents/organization/250044.pdf

Though certainly an unprecedented advancement in HIV research, PrEP use has also sparked concerns from individuals questioning its safety. In a New York Times column (2012) Denise Grady outlines potential risks associated with, "a drug that *healthy* people take once a day to prevent HIV infection" (p.D5, emphasis added). Her concerns stem from a potentially misleading logic that an oral pill is the equivalent of other proven safe sex practices, such as condom use. Costa-Roberts (2015) further addresses the often undiscussed long-term adverse effects of daily PrEP uptake among *healthy* adults, including: "lactic acidosis, …liver problems, kidney issues — including

[23] The Determined, Resilient, Empowered, AIDS-free, Mentored and Safe women (DREAMS) initiative is a partnership with PEPFAR intended to reduce HIV infection in ten Sub-Saharan African countries(Chasela et al., 2010).

kidney failure — and bone density loss" (Retrieved from: https://www.pbs.org/newshour/health/8-things-didnt-know-truvadaprep).

Concerns over PrEP's safety came under more scrutiny after two lawsuits were filed against Gilead Sciences, the company that manufactures Truvada, due to adverse effects from continued, exposure to Truvada (Peterson, 2018). The lawsuits claim:

> …instead of continuing to develop a safer alternative, [Gilead Sciences] decided to hide tenofovir's risks [bone density loss and kidney failure] while earning billions of dollars as it became one of the world's most prescriptible medicines for HIV (retrieved from: http://www.latimes.com/business/la-fi-gilead-hiv-drug-lawsuit-20180509-story.html).

Therefore, the timeliness of this example can serve as a strong heuristic tool with which to test the application of GRADEpro. Employing the GRADE framework for evaluating investigations studying PrEP will help determine if GRADE (a) serves as a tool for detecting bias and (b) can be adopted more commonly as a novel tool with which to detect number bias.

Methods

Sample

To exemplify the use of GRADEpro for evaluating the quality of evidence generated by large clinical trials I utilized the four investigations upon which the FDA

grounded its approval of PrEP for prophylactic use[24]: (1) Preexposure Chemoprophylaxis for HIV Prevention in Men who Have Sex with Men, also known as 'iPrEX' (Grant, et al., 2010), (2) Antiretroviral Preexposure Prophylaxis for Heterosexual HIV Transmission in Botswana, or 'TDF-2', for short (Thigpen, et al., 2012), (3) Antiretroviral Prophylaxis for HIV Prevention in Men and Women, referred to as 'Partner's PrEP' (Baeten, et al., 2012), and (4) Antiretroviral Prophylaxis for HIV Infection in Injecting Drug Users in Bangkok, Thailand, or the 'Bangkok Tenofovir' study (Choopanya, et al., 2013).

These studies were selected for analysis for several reasons. First, each clinical investigation is considered 'pioneer' research for PrEP. In other words, these investigations were the first to support, and eventually endorse the approval of Truvada as a prophylactic drug by the Food and Drug Administration. Further, despite the availability of numerous studies on PrEP's effectiveness among various populations, the four aforementioned studies were the studies showcased as evidence for PrEP on the CDC website.

Second, the studies were planned and conducted as Randomized Control Trials (RCT), often considered the most valid design for clinical research (Bothwell, Greene, Podolsky, & Jones, 2016). It is important to note, however, that despite being touted as the 'gold standard', RCTs are currently under increased scrutiny for potentially biased

---

[24] The FDA's statement can be found on the Gilead Sciences home page under 'Press Releases' at: http://www.gilead.com/news/press-releases/2012/7/us-food-and-drug-administration-approves-gileads-truvada-for-reducing-the-risk-of-acquiring-hiv

applications (Rastrollo, Schulze, Ruiz-Canela, & Martinez-Gonzalez, 2013; Chan & Altman, 2005; Chopra, 2003). Therefore, PrEP is a strong heuristic example with which to test GRADEpro as a valid measure to parse out important information on the quality of PrEP's supporting research.

Analysis

Once identified, the studies underwent a two-step analytic process involving: (1) descriptive analysis of the presentation of findings, and (2) application of GRADEproGDT software to rate evidence quality. The purpose of using a two-step analysis was to help generate a well-rounded understanding of the quality of the investigations' findings. Specifically, rating the quality of evidence solely via GRADEpro alone would only provide information on evidence quality and *not* a thorough report on common reporting practices. Conversely, discussing only the number of *p*-values relative to effect sizes and confidence intervals might not paint a comprehensive and accurate picture of the evidence's quality.

Analysis - Step 1: Descriptive Analysis. The descriptive portion of this analysis was used to answer the question: *What are some of the common reporting patterns in this sample of PrEP clinical trials regarding: (a) efficacy, and (b) side effects?* The purpose of this question was to assess common traits and patterns used to relay information to readers. To answer this question, I extracted and examined pertinent information related to: (1) the number of significance tests in each study, (2) the proportion of significant to non-significant findings, and (3) the use of other measures to convey importance of findings (e.g. confidence intervals or effect sizes).

Among other factors, these data can highlight the degree to which certain practices are favored over others (i.e. how many *p*-values versus confidence intervals, types of effect sizes, among others).

The rationale for extracting this information from each of the studies assessed is as follows. First, Lambdin (2012) contends the nature of significance testing can only dichotomously inform researchers about the non/significance of a finding. In other words, the *p*-value tests the null/nill hypothesis that the effect studied is equal to zero. Therefore, a significance tests can only inform whether the effect is equal to zero or not equal to zero— but *not* how far away from zero (or strong) the effect is (Thompson, 2003). Thus, what is presented as a significant finding is *less informative* than other statistics such as confidence intervals and effect sizes.

Even though *p*-values relay little information, their ease of interpretation make them popular in today's academic climate and they are often viewed as the academic gold standard (Nuzzo, 2014). Continued reliance on significance tests, however, downplays the role of other important measures, such as confidence intervals and effect sizes, which are better indicators of overall effectiveness (Lee, 2016). Therefore, in today's academic climate we have a collection of literature, reliant on p-values, that fails to pay adequate attention to *what* those significant and non-significant findings *mean*.

Analysis – Step 2: The Application of GRADEpro. For the second step in the analysis, I utilized the Cochrane Institute's GRADEproGDT (otherwise known as GRADEpro) software to answer a second question: *What is the quality of the numeric*

evidence regarding: (a) the efficacy of PrEP, and (b) its reported side effects[25]*?* I purposefully elected to use GRADEpro over other open source software packages, such as "statcheck", for two reasons. First, the algorithm used in statcheck can only identify study findings if they are formatted in the American Psychological Association (APA) format. The articles reporting on the PrEP clinical trials were formatted following the American Medical Association standards (AMA) and, thus, statcheck could not recognize the data. Second, when compared to statcheck, the GRADEpro software is more commonly used and its testing and validation have been reported through publications authored by the GRADE Working Group— an international, open collaboration dedicated to promoting transparency in research— and in other systematic reviews implementing the software (Kavanagh, 2009; Puhan et al., 2014; Zhao, Liang, Fang, & Liu, 2017).

GRADEpro's software interface allows users to create an Evidence Table from an aggregate collection of numeric data. Each evidence table consists of two components: (1) the assessment of quality, and (2) the summary of findings. The assessment of quality section evaluates five dimensions of the *overall* study, qualitatively: (a) Risk of Bias, (b) Imprecision, (c) Inconsistency, (d) Indirectness, and (e) Publication Bias. The summary of findings table determines overall precision of *individually* tested hypotheses *within a study* by (a) extrapolating results per hypothesis

[25] The rationale for selecting efficacy and side-effects as the two testable hypotheses was (1) the generally touted effectiveness of PrEP, and (2) concerns about long-term side effects due to daily dosing.

(e.g. number of events that occurred within treatment and control groups, the Hazard Ratio [HR], and the confidence interval about that HR).

Analysis – Step 2 (a): Assessment of Quality. Because GRADEpro evaluates the inputted results systematically, results derived from this analysis will elucidate the level of evidence *quality* among five key dimensions: (1) Risk of Bias, (2) Imprecision, (3) Inconsistency, (4) Indirectness, and (5) Publication Bias. The evaluator then determines the extent of 'risk' associated with each dimension, by evaluating potential infractions as: (a) not serious, (b) serious, and (c) very serious (see Appendix B).

Per Cochrane's definition, dimension one, risk of bias, refers to any internal or external factor that could influence a study's results. Internally (i.e. pertaining to study design), the GRADE working group has a series of preferences regarding how a study should be designed. For the working group, studies are at no risk of bias if they use randomized, double-blind, controlled trials. Studies can lose up to two points, if a study deviates from that standard— (e.g. does not blind, does not use RCT design, does not disclose methodology). External risk of bias refers to factors that can sway findings (e.g. attrition, ending the study early, source of funding, among others).

Indirectness refers to the extent to which the sample used in the clinical trial reflects the population for which the drug/intervention is intended. In other words, the medication should be tested among the most likely users, only. Imprecision, refers to the 'tightness' of findings. Studies with congruent, homogenous findings are upgraded, while sporadic findings are downgraded— one recommendation by Cochrane is to look

at confidence intervals. Large confidence intervals are indicative of sporadic findings that mar interpretation of the data.

Inconsistency refers to the research question matching the analysis— i.e., did the study measure what it intended to measure? Publication Bias is a category designed to capture the evaluator's subjective suspicions of publication bias. Studies interpreted as exhibiting publication bias are downgraded. The last criterion, dose-response gradient, can help a study recover one deducted point if a gradient is included in the study— a dose response gradient is a chart, or graph, the measures the least amount of medication necessary to have an effect.

After each of the five dimensions receives a quality score, GRADEpro aggregates those scores to calculate overall level of evidence quality. Those levels are: (1) High: There is ample evidence the true effect lies close to the estimated effect, (2) Moderate: There is modest evidence the estimated effect lies close to the true effect, but there is some possibility the true effect is different, (3) Low: There is little evidence to support the estimated effect reflects the true effect, and (4) Very Low: There is limited or no evidence to suggest the estimated effect is remotely near the true effect.

Analysis – Step 2 (b): Summary of Findings. The summary of findings table uses information from tested hypotheses in each study to calculate an absolute confidence interval. The purpose of the absolute confidence interval is to put a relatively misunderstood statistic, Hazard Ratio (Spruance, Reid, Grace, & Samore, 2004), into a statistic that lay readers can better interpret. For example, in a double blind 1:1 randomized control trial testing the efficacy of a medication, the evaluator

would extrapolate the number of events— infections— within the total sample for both treatment and control groups.

The number of events, total sample, and corresponding HR and CI would then be entered into GRADEpro. The program then uses that data to calculate what Cochrane calls an "absolute confidence interval", which presents those findings in terms of number of cases per 'x' events. In other words, within our current example, rather than interpret the following HR: .56, 95% CI: [.22-.68], the absolute CI is phrased as: *per 1,000 cases there would theoretically be 150 fewer instances of infection, with a range of 110-180.*

Results

1. Step 1 - Descriptive Analysis: ***What are some of the common reporting patterns in the PrEP clinical trials regarding (a) efficacy and (b) side effects?***

For the descriptive analysis on both the efficacy and side-effects hypotheses, the four clinical trials were evaluated on the following criteria: (1) the prevalence of significance tests for each hypothesis-- (a) overall efficacy, and (b) side effects; (2) the proportion of significant to non-significant findings; (3) the use of confidence intervals and effect sizes, and (4) sample size of the overall study.

With regard to efficacy (i.e. the main hypothesis of the studies), there were a total of 12 significance tests among the four clinical trials (see Table 3.1). Of the 12 significance tests, approximately 95% (n = 11) were statistically significant at $p<.05$. Each significance test had one corresponding effect size, HR, and a corresponding

interval around each HR. The sample sizes were as follows: (1) iPrEX n= 1,248, (2) TDF-2 n= 1,219, (3) Partner's PrEP n=1,534, and (4) Bangkok Tenofovir n= 4,823.

For side effects there was a combined total of 74 reported *p*-values testing differences among treatment and control groups over a variety of side effects (e.g. nausea, headache, diarrhea, vomiting, among others). Of those *p*-values, approximately 85% (n=60) were non-significant at the .05 level of probability, and 15% (n=14) were significant. None of the studies used other metrics to substantiate findings— such as confidence intervals or effect sizes.

2. Step 2 - *The Application of GRADEpro*

The GRADE approach to evaluation is unique because it is divided into two separate components: (1) assessment of quality, and (2) summary of findings. Using the GRADE approach, then, allows one to evaluate the study overall (assessment of quality) and each individual hypothesis (summary of findings). Below, I assess the quality, overall, and in the summary of findings section I evaluate two hypotheses: (1) efficacy and (2) side effects.

For the assessment of quality, GRADEpro prompts users to evaluate studies on five dimensions: (1) Risk of Bias, (2) Imprecision, (3) Inconsistency, (4) Indirectness, and (5) Publication Bias. For each dimension, there are three options on a point-and-click menu interface the evaluator can choose: (1) not serious, (2) serious, and (3) very serious. Evaluators employ the Cochrane criteria for 'upgrading' or 'downgrading' studies by making value judgements when determining the severity of infraction of those domains (see Appendix B for details on the grading system/criteria).

In the summary of findings table, evaluators must input information on the selected hypotheses into the program, manually. In other words, comparisons between treatment and control, the corresponding *p*-value, and any CIs or HRs must be typed into the appropriate boxes. From that information, GRADEpro produces the absolute confidence interval, which interprets HRs, RRs, and ORs in a different language. For example, rather than say HR: .54, 95% CI [.22-.88], the absolute CI would read: "For every 1,000 cases, there were 'x' fewer cases ranging from 'y' fewer cases to 'z' fewer."

Step 2 (a) - Assessment of Quality. Per Cochrane's evaluation criteria, all studies began with a grade of 'high' because each one employed a RCT design (refer to Appendix B for a detailed description of Cochrane's evaluation criteria). Each study, however, was downgraded by one or two points along the evaluation process resulting in the final grades: (1) High—iPrEX, (2) Moderate— Partner's PrEP & Bangkok Tenofovir, and (3) Very Low— TDF2 (See Table 3.2).

For risk of bias, both iPrEX and TDF-2 were downgraded one point each, but for different reasons. IPrEX disclosed several of its investigators received financial compensation from Gilead, the manufacturer of Truvada, while the TDF-2 study was terminated early due to large attrition rates. Therefore, the researchers did not report any findings on efficacy. Further, because TDF-2 did not present general findings or conclusions on the efficacy of PrEP, that study also was downgraded on inconsistency (defined as the testing of hypotheses in a similar manner).

Studies are *indirect*, according to Cochrane, when they are conducted on samples that are not the intended population. Therefore, because TDF-2, Partner's PrEP and Bangkok Tenofovir did not include US participants— and PrEP is often viewed as a US-centered medication—those studies were downgraded one point. IPrEX, however, did not lose points because it included a small sub-sample of US participants (n=247) as part of a larger, international sample (n=2499).

Three of the four studies, iPrEX excluded, were downgraded one point on imprecision into the 'serious' category, because of concerns with large confidence intervals marring the interpretability of any hypothesis. In TDF-2, Partner's PrEP, and Bangkok Tenofovir, there were multiple examples in which the confidence intervals throughout the document, as part of small sub-tests within the grand efficacy hypothesis, were notably large and reported without any accompanying explanation (e.g., iPrEX— "the odds of HIV infection were lower by a factor of 12.9 (95% CI, 1.7 to 99.3; $p<0.001$"; TDF-2— "the protective efficacy was 61.7% (95% CI, 15.9 to 82.6 $p=0.03$; and Partner's PrEP— there was a 75% relative reduction due to Truvada, "95% CI, 55 to 87, $p<.0.001$").

Upon completing the evaluation, and per Cochrane's criteria, studies are to be awarded one additional point if they reported a dose-response gradient— consisting of a chart that determines lowest dosage level for needed effectiveness of treatment drugs. All studies, except for TDF-2, reported a dose-response gradient and, therefore, were awarded the extra point.

Step 2 (b) - Summary of Findings

PrEP Efficacy. The summary of findings for the overall efficacy hypothesis (see Table 3.4) shows that, in each study, regardless of its final grade, there was a marked reduction in the number of HIV infections among the treatment group. Each study, further, reported a Hazard Ratio (HR), and CI, to substantiate findings. The corresponding HRs for efficacy were: iPrEX— HR: .53 (.36 to .85); TDF-2— N/A; Partner's PrEP – HR: .27(.16-.48); and Bangkok Tenofovir— HR: .35 (.21 to .56). (See Table 3.3).

Because the studies reported HRs and CIs to substantiate findings, when that information is typed into GRADEpro, GRADEpro calculates an absolute confidence interval for each study. An absolute confidence interval is, generally, used in meta-analysis to aggregate various HRs and CIs into a general statistic intended to make sense (i.e. interpret) of what all HRs mean as an aggregate. However, the absolute CI can also be used for individual HRs and CIs, to make HR easier to interpret.

The absolute CI for iPrEX, the study reporting an HR of .53 (.36 to .85), is 27 fewer cases per 1,000 with a range of 8 to 37 fewer cases. TDF-2, again, did not report any findings; therefore, I could not calculate the absolute CI. Partner's PrEP (TDF Only) reported an HR of .27(.16-.48), which translates to an absolute CI of 23 fewer cases per 1,000 ranging from 19 to 31 fewer. Parter's PrEP (TDF-2), HR .27 (.16 to .48), translates to 27 fewer cases per 1,000, ranging from 19 to 31 fewer. Finally, Bangkok's Tenfovir study reported HR for efficacy as .35 (.21 to .56), translating to an absolute CI of 4 fewer cases for 1,000, ranging from 3 fewer cases to 5 fewer cases.

Side Effects. A second summary of findings was also created for the following hypothesis—*There are no statistical differences between treatment and control groups in observed increases in creatinine levels*[26]. Creatinine— which is a kidney-damaging chemical waste molecule generated by muscle metabolism and measured via blood tests (Davis, 2017) — was the side effect considered for analysis, here, because it was the only common serious side effect found in each of the four studies. For this hypothesis, the Hazard Ratios for each study were calculated by hand[27], as none of the studies reported HRs when presenting the findings related to adverse events (see Table 3.4).

TDF-2, iPrEX, Partner's PrEP, and Bangkok Tenofovir all reported non-significant differences ($p > .05$) between treatment and control groups in observed increases of creatinine blood-levels. The *p*-value was the only statistic used to relay information. None of the studies reported a corresponding HR and CI for this hypothesis. Therefore, the HRs presented in Table 3.5 were calculated by hand, to allow GRADEpro to provide an absolute CI.

When comparing iPrEX's treatment and placebo groups, there were 13 more people with increased creatinine, representing a HR of 1.96 and an absolute CI of 5

---

[26] Creatinine is measured via a common blood test. A healthy amount of creatinine in the blood can range from .87 to 1.2. Study protocols for the 4 studies indicated using DAIDS AE grading table as set criteria to determine the severity of increased creatinine among patients in their respective samples.

[27] Because HR was calculated by hand and there is no other information about the data, such as standard deviations, there is no confidence parameter for the HR's associated with increased creatinine.

more cases per 1,000. In the TDF-2 study, there was 1 fewer documented case when comparing treatment and placebo groups (HR .95; Absolute CI 1 fewer per 1,000 cases). Bangkok Tenofovir, had 85 more cases of increased creatinine in the blood (HR 3.36; absolute CI 67 more cases per 1,000).

Discussion

1. Descriptive Analysis: ***What are some of the common reporting patterns in the PrEP clinical trials regarding (a) efficacy and (b) side effects?***

The reporting practices in the PrEP clinical trials exhibit some of the problematic patterns documented previously by several scholars (Thompson, 1999; Ioannidis, 2005; Nuzzo, 2013; Hubbard, 2015). Two patterns, however, stand out in the presentation of findings in each report: (1) overreliance on *p*-values and (2) selective reporting of CIs and effect sizes.

Both hypotheses in each study (i.e., efficacy and side effects) primarily presented findings via *p*-values— findings for efficacy were reported with 11 significant and 1 non-significant test; side effects, with 14 significant and 60 non-significant tests. Based on the criteria of statistical significance, these findings support the hypothesis that PrEP is more effective than a placebo for preventing HIV infections among people at risk, with low risk of side-effects.

Though using *p*-values to report effectiveness is not *necessarily* problematic, it is important to qualify that the studies assessed here relied on large sample sizes: 1,248 for iPrEX; 1,219 for TDF-2, 1,534 for Partner's PrEP, and 4,823 for Bangkok Tenofovir. As Thompson (2003) contends, *p*-values have a unique mathematical

relationship with sample size, in which, considering the analysis employed, will almost always return a significant *p*-value.

To mitigate *p*-value bias associated with large sample sizes, the American Psychological Association Task Force for Statistical Inferencing (1999) recommends scholars report effect sizes and confidence intervals to assist in the interpretation of a significant *p*-value. Though the PrEP clinical trials did report CIs and effect sizes— in this case Hazard Ratios (or, number of events/infections in the control group divided by events in the treatment group)— they were reported selectively. In fact, *p*-values were the *primary* metric for reporting findings, and they were reported twice as frequently than other metrics for the side-effects hypotheses.

For efficacy, the results were presented with a *p*-value, HR, and corresponding confidence interval for the HR. Though none of the studies *interpreted* the HR or CI, these metrics still represented more information than what was provided for the testing of the side effects hypotheses. Treatment and control comparisons for side effects were *only* reported with a *p*-value, in each study. There was no reporting of CIs or HRs to contextualize and substantiate findings. Further, there were no other descriptive data for the side-effects hypothesis— such as means or standard deviations—that would allow readers/evaluators to calculate appropriate CIs. Therefore, it was not possible to calculate confidence intervals from the data provided.

The selective reporting of CIs is problematic. In three of the four studies, with regard to increased creatinine levels, specifically, there were more cases in the treatment group than the control group. Though some of the numbers of patients

experiencing increased levels were small— e.g. iPrEX 5 additional cases per 1,000— the Bangkok Tenofovir's 67 more cases per 1,000, with a HR of 3.36, is alarming, despite the non-statistical-significance of the finding. However, as Thompson (2002) argues,

> Statistical significance [and non-significance] is not sufficiently useful to be invoked as the sole criterion for evaluating noteworthiness in research…[we]…should expect a literature in which the results of a single study are explicitly interpreted using effect sizes in direct comparisons with the typical effect sizes from previous studies (p.66).

2. *The Application of GRADEpro: What is the quality of numeric evidence regarding: (a) the efficacy of PrEP, and (b) side effects*?

Assessment of Quality. Based on Cochrane's evaluation criteria, the PrEP clinical trials raise two important concerns: indirectness and imprecision. Indirectness, specifically, refers to concerns regarding the nation in which a study was conducted. Imprecision, on the other hand, refers to large gaps in confidence intervals, which taint the interpretation of a tested hypothesis. In other words, during the qualitative assessment of each study, I determined, via Cochrane's evaluation sheet, if infractions within each category were severe enough to downgrade the quality of each study (downgrading meant deducting one point, and lowering the overall grade).

With regard to indirectness, Shunemann (2011) states studies should be downgraded at least one point (into the "serious" category) if a medication intended for use in *low income* countries is tested exclusively with samples from high income

countries and vice-versa. In other words, the medication should be tested with the population the medication intends to treat.

Of the four PrEP clinical trials, only one study, iPrEX, relied on a sub-sample of US participants as part of a larger, international sample. The remaining studies, TDF-2, Partner's PrEP, and Bangkok Tenofovir selected study sites, and used a sample of individuals, from the following countries: Thailand, Botswana, Kenya, and Uganda, respectively. Those studies did not sample from a US population.

Though the practice of selecting samples from other nations may not seem problematic, by Cochrane's standards, studies that do so are considered "indirect". In the case of PrEP, this finding is especially important because PrEP is viewed as a US-centered phenomenon. The website www.prepwatch.org (2018), for example, recently published global PrEP uptake rates. These rates reveal the United States is the leading nation in number of PrEP prescriptions, at an estimated 220,000-250,000 unique prescriptions written for Truvada, to date. These numbers stand in sharp contrast to the prescriptions written in countries where the original trials were conducted— (1) Thailand, (4,000-5,000), Botswana (0-200), Uganda (4,000-5,000), and Kenya (25,000-26,000). Even in other, more affluent nations, numbers on PrEP uptake remain similarly low— England (4,500-55,000), China (500-700), and Canada (900-1,100).

Imprecision was a second major factor affecting the overall quality rating (defined by Cochrane as large confidence intervals marring the interpretation of the data). I elected to downgrade studies one point (into the "serious" bias risk category) because many reported confidence intervals from sub-tests (those other than the overall

hypothesis for efficacy) were notably large and unexplained. For example along with large CI's mentioned previously, in TDF-2 "the overall protective efficacy of TDF–FTC in the modified intention-to-treat analysis (comprising 1,216 participants) was 62.2% (95% CI, 21.5 to 83.4; p=0.03)[28]" (Thigpen et al., 2122)". Further, in Partner's PrEP, "reductions in the rates of HIV-1 acquisition of 67% due to [Tenofovir] (95% confidence interval [CI], 44 to 81; P<0.001) and 75% due to Truvada (95% CI, 55 to 87; P<0.001)" (Baeteen et al., 2012, p. 1223).

The problem with large confidence intervals, contends Thompson (1999), is that one cannot have as much faith in the point estimate as one would with narrow confidence intervals. In other words, CIs with a narrower range between upper and lower bound limits infer a more stable treatment effect. Therefore, contends Field (2011), one should also (1) rely on other effect sizes to convey meaning, and (2) explain the reported statistics. Failure to do so will prompt readers to come to a potentially incorrect conclusion.

Another issue, present in two of the studies (iPrEX and TDF-2), was potential risk of bias in the risk of bias domain. The risk of bias domain, according to Cochrane's evaluation criteria, is defined as any systematic or outside influences that sway the credibility of the evidence (Cochrane 2013). Two of the studies were rated as "serious risk" due to questionable funding relationships: The TDF-2 and iPrEX studies.

---

[28] Though TDF-2 reported this finding in-text, they did not report the frequency of infection, along with corresponding HR and CI, as did the other studies. This was not done because the study terminated early and could make conclusions.

Chopra (2003) contends one of the strongest promoters of bias in clinical research is funding source— particularly when that funding stems from industry or corporations. Though all four clinical studies obtained the tested medications, free of charge, from Gilead Sciences (the pharmaceutical company manufacturing Truvada), the iPrEX and TDF-2 studies disclosed information in their respective acknowledgement sections that warranted further concern and evaluation. The iPrEX authors, for example, disclosed financial support from Gilead beyond the study medication. Specifically, some investigators were currently funded by Gilead via unrelated grant mechanisms, such as support for studies other than PrEP. Others disclosed being on Gilead's payroll, or holding stock in the company:

> Dr. Mayer reports receiving grant support from Gilead, Merck, and Bristol Myers Squibb; Dr. Kallás reports serving on a data and safety monitoring board for Merck; Dr. Schechter reports receiving consulting fees and grants from Gilead; Drs. Liu and Anderson report receiving donations of study drug from Gilead for various PrEP projects; and Drs. Jaffe and Rooney report being employees of Gilead Sciences and owning stock in the company. No other potential conflict of interest relevant to this article was reported (iPrEX; Grant et al., 2010, pg. 2598).

Though certainly not to the extent of iPrEX, the authors of the TDF-2 study disclosed distinct information, as well, suggesting that members of the team had prior exchanges with Gilead, beyond the scope of the clinical trial:

> Dr. Hart reports receiving royalties from Roche Diagnostics; the agreement with Roche pays the Centers for Disease Control and Prevention and project investigators annually for the rights to use a molecular clone in the test kits. Dr. Hendrix reports receiving grant support from Gilean (*sic*) Sciences (Thigpen et al., 2012, p.433).

The remaining two investigations' (Partner's PrEP and Bangkok Tenofovir) interactions with Gilead were limited to receiving the study medication as a donation.

Summary of Findings: Efficacy. The CDC (2017) touts, "Daily PrEP reduces the risk of getting HIV from sex by more than 90%. Among people who inject drugs, it reduces the risk by more than 70%" (retrieved from: https://www.cdc.gov/hiv/basics/prep.html). All four studies listed on the CDC website support this statement— PrEP is more effective than a placebo at preventing the \ infection by HIV in each of the following groups: (1) men who have sex with men (iPrEX), (2) heterosexual serodiscordant couples (TDF-2), (3) active drug users (Bangkok Tenofovir), and (4) heterosexual men and women (Partner's PrEP).

To support the findings related to efficacy, the investigators appropriately reported a corresponding HR— but HRs were reported *only* for the efficacy hypotheses. Authors failed to report an HR for the other important hypotheses tested, such as side effects. However, regardless of the level of significance, researchers should be more concerned with the size of the effect (Thompson, 2002). More often than not, a non-significant finding will have some treatment effect that warrants concern or further evaluation (Thompson, 1999).

GRADEpro uses the HRs provided in each study to calculate an absolute confidence interval for that HR — which phrases the imputed HR in more interpretable language. All absolute confidence intervals calculated by GRADEpro comprised larger deviations from the point estimate than the original HRs, suggesting the studies' effects might be smaller than perceived. The iPrEX study reported an HR of .53, which translated into27 fewer cases per 1,000 with a range of 8 fewer to 37 fewer. For the Bangkok Tenofovir's study, the HR of .35 translated into 4 fewer cases per 1,000, with a range of 3 fewer to 5 fewer. Thus, while the medication appears to prevent infection in the experimental groups at levels that exceed chance, the HR effect size allowed for a more accurate picture. Coupled with the corresponding absolute Cis, the resulting image becomes even clearer.

Though each study's absolute CI and HR indicated PrEP was *more effective* than a placebo for the groups studied, it is worth noting there were at least two other trials — conducted around the same time and funded by similar grant mechanisms — not listed on the CDC website: FEMPrEP and Vaginal and Oral Interventions to Control the Epidemic (VOICE) (See Figure 3.1). These two studies, published in the New England Journal of Medicine, tested the efficacy of PrEP among heterosexual women in Africa. Unlike the CDC-listed trials, however, these studies found *no* significant differences between treatment and control groups in PrEP's efficacy (See Table 3.5).

Because FEMPrEP and VOICE completed their respective trials and included overall efficacy results, I elected to calculate an absolute CI from the HRs reported for

the overall efficacy hypothesis tested in each study. In both studies, more people were infected in the treatment group resulting in an absolute CI highlighting the ineffectiveness of Truvada among those samples. Specifically: FEM-PREP — HR .94 (.59 to 1.52), or 2 fewer cases per 1,000 (from14 fewer to 17 *more*) and VOICE Tenofovir— HR 1.49 (.97 to 2.29) 20 more cases per 1,000 (from1 fewer to 51 *more*). Despite the documented ineffectiveness, however, PEPFAR still recommends PrEP uptake for women in Africa and for female sex workers.

Though the omission of FEMPREP and VOICE from the CDC website is certainly grounds for downgrading the other four studies into the 'serious' category (due to selectivity bias), I elected to not do so, as there is no evidence these studies were intentionally omitted. However, acknowledging their existence, and their contradictory findings, is important when promoting PrEP.

Summary of Findings: Side Effects. Despite all studies reporting a list of mild and moderate side-effects, each investigation relied exclusively on *p*-values to report findings. HRs were not reported for side-effects, nor was the statistical test used to arrive at the significant/non-significant findings disclosed. Though the studies listed many side-effects (nausea, headache, fatigue, among others), increases in creatinine blood-levels were reported in all four studies. This side-effect raises concerns, because increases in creatinine can lead to renal failure (Herget-Rosenthall et al., 2004). It is

worthy of noting that renal failure is one of the side-effects at the center of a recent class-action lawsuit filed against Gilead Sciences[29] in the US.

The frequencies of increased creatinine, alone, show that, in most studies, there were more cases of increased creatinine in the treatment group, than in the control group. By calculating the HR for those frequencies, and then calculating an absolute CI from those HRs, we arrive at generalized findings that examine how many more cases of creatinine we would expect per 1,000 cases. As with the efficacy hypothesis, the HR and absolute CIs for the side-effects varied. Specifically, iPrEX had 5 more cases of increased creatinine per 1,000; TDF-2, was not estimable; Partner's PrEP revealed 1 *fewer* per 1,000; and Bangkok Tenofovir 67 more per 1,000.

These analyses indicate that reliance on statistical significance testing, alone, can obscure important clinical findings. While increased creatinine levels in PrEP patients is documented and accepted within the medical community as a side-effect, the actual kidney effects and damage are often ignored in discussions of PrEP. Such dismissal is grounded in part, in the findings reviewed here. Most health care providers take at face value the claim that the "PrEP clinical trials did not find significant negative effects on kidney function" (Highleyman, 2018, retrieved at: https://betablog.org/new-research-at-croi-2016-how-pre*p*-changes-kidney-function/).

---

[29] "Two Southern California men filed suit against Gilead Sciences [in May 2018] …The lawsuit says that HIV positive patients suffered from as many as 10 years of additional accumulated kidney and bone toxicity while using the drug…sold under…brand names, including Atripla, Truvada, Stribilb and Complera" (Peterson, 2018, retrieved from: http://www.latimes.com/business/la-fi-gilead-hiv-drug-lawsuit-20180509-story.html).

However, as more cases of creatinine rising-levels and renal damage occur, the issue is gaining attention. At the 2018 Conference on Retroviruses and Opportunistic Infections (CROI), for instance, researchers convened a panel dedicated to discussing the 'modest' kidney changes associated 'mild' kidney damage resulting from PrEP uptake (CROI, 2018).

However, with no effect size presented in the studies for this hypothesis, we do not have an accurate perspective. As evidenced by the externally calculated absolute confidence intervals, even non-significant *p*-value provide valuable information; they should, therefore, never be ignored and always be reported (Thompson, 2003).

Conclusion

In this chapter, I argued for the importance of detecting numeric bias in academic scientific work, and demonstrated the utility of employing a tool such as GRADEpro to help identify potential numeric biases in research reporting. To accomplish that task, I opted to use a modern, heuristic example of a medication that is currently being promoted (via policy and activism) and evaluate seminal studies of that medication using GRADEpro. The evaluation comprised two components: (1) a descriptive analysis, and (2) a GRADEpro evaluation.

While the descriptive analysis yielded important and insightful information, it was the GRADEpro tool that provided a well-rounded and thorough evaluation of each study on five domains: (1) Risk of Bias, (2) Imprecision, (3) Inconsistency, (4) Indirectness, (5) and Publication Bias. Using the Cochrane Evaluation criteria to determine if mild or severe infractions occurred in those domains, studies could receive

one of the following GRADES: (1) High (strength of evidence), (2) Moderate, (3) Low, and (4) Very Low.

For various reasons, studies included in the heuristic example were downgraded on one or more issues among the five domains. Thus, GRADEpro, and the criteria used to evaluate studies, successfully--and more objectively-- assisted in identifying problematic issues: (a) within study execution and (b) other suspect situational factors potentially influencing the outcome of one, or more, investigations. Further, GRADEpro's ability to calculate an absolute CI from reported data proved useful to help further interpret the effect size used in the analyzed studies— Hazard Ratios. The information derived from the absolute confidence interval added further context, and meaning, to otherwise un-discussed effect sizes.

Despite its utility, it is important to bear in mind the inherent limitation in the GRADEpro tool. While the tool is more objective that current measures, such as self-evaluation, interpretation of the Cochrane criteria and decisions on to downgrade for various infractions ultimately rests with the researcher. Therefore, we cannot be one hundred percent certain our own biases influenced the decision to downgrade studies into the 'serious' or 'very serious' categories.

Overall, GRADEpro did what researchers at Cochrane intended— provide an accessible, open source software platform with which to assess the quality of large-scale investigations and promote transparency in research. Results from this example further highlight the importance, and utility, of programs such as GRADEpro for future efforts in numeric bias detection and mitigation.

Table 3.1. The number of total reported significant/non-significant *p*-values, confidence intervals, effect sizes, and sample sizes for each of the four PrEP clinical trials testing the overall efficacy hypothesis.

	Significant p-values	Non Significant p-values	Total p-values	Confidence Intervals	Effect Sizes	Sample Size
iPrEX	3	1	4	4	4	1248
TDF-2	-	-	-	-	-	1219
Partner's PrEP	4	-	4	4	4	1534
Bangkok Tenofovir	4	-	4	4	4	4823
Total	11	1	12	12	12	8824

Table 3.2. Assessment of Quality using the Cochrane Criteria to determine the overall strength of evidence of 4 clinical trials studying PrEP's efficacy.

	Study Design	Risk of Bias	Inconsistency	Indirectness	Imprecision	Grade
iPrEX (TDF-FTC)	RCT	Serious	Not-Serious	Not-Serious	Not-Serious	**** (High)
TDF-2 (TDF-FTC)	RCT	Serious	Serious	Serious	Serious	* (Very Low)
Partner's PrEP (TDF-FTC)*	RCT	Not-Serious	Not-Serious	Serious	Serious	***(Moderate)
Partner's PrEP (TDF)*	RCT	Not-Serious	Not-Serious	Serious	Serious	***(Moderate)
Bangkok Tenofovir (TDF-FTC)	RCT	Not-Serious	Not-Serious	Serious	Serious	***(Moderate)

* Partner's PrEP has two rows because they used a 1:1:1 design, which tested (1) Tenofovir versus Placebo, and (2) Truvada versus placebo.

Table 3.3. Summary of Findings for the efficacy hypothesis tested in 4 clinical trials of PrEP, examining: (1) the number of HIV-infections among treatment and control groups, the corresponding effect size (Hazard Ratio) and (2) a computer-calculated absolute confidence interval.

	Number of Patients		Effect	
	PrEP	Placebo	Relative	Absolute
iPrEX (TDF-FTC)	38/1251	72/1248	HR .53 (.36 to .85)	27 fewer cases per 1,000 (from 8 to 37 fewer)
TDF-2 (TDF-FTC)	-	-	-	-
Partner's PrEP (TDF-FTC)	22/1584	58/1584	HR .38 (.23 to .62)	23 fewer cases per 1,000 (from 14 fewer to 28 fewer)
Partner's PrEP (TDF)	16/1579	58/1584	HR .27 (.16 to .48)	27 fewer cases per 1,000 (from19 to 31 fewer
Bangkok Tenofovir (TDF-FTC)	17/4843	33/4823	HR .35 (.21 to .56)	4 fewer cases per 1,000 (from 3 to 5 fewer

Table 3.4. Summary of findings regarding PrEP side-effects comparing observed blood levels of creatinine (> 1.2 per mm) between treatment and control groups in four clinical trials.

	Number of Patients		Effect		*p*-value
	PrEP	Placebo	Relative	Absolute	
iPrEX (TDF-FTC)	28/1251	15/2611	HR 1.86	5 more cases per 1000	0.08
TDF-2 (TDF-FTC)	-	-	-	-	-
Partner's PrEP	19/1584	20/1579	HR .95	1 fewer case per 1,000	0.28
Bangkok Tenofovir (TDF-FTC)	121/1204	36/1209	HR 3.36	67 more cases per 1,000	0.09

Table 3.5. Summary of Findings Table for two additional studies of PrEP's efficacy -- not listed on the CDC website.

	Number of Patients		Effect	
	PrEP	Placebo	Hazard Ratio	Absolute CI
FEMPReP	33/1024	35/1032	HR .94 (.59 to 1.52)	2 fewer cases per 1,000 (from14 fewer to 17 more)
VOICE TDF	52/823	35/838	HR 1.49 (.97 to 2.29)	20 more cases per 1,000 (from1 fewer to 51 more)
VOICE TDF-FTC	61/1284	60/1308	HR 1.04 (.73 to1.49)	2 more cases per 1,000 (from 12 fewer to 22 more)

Figure 3.1. A screenshot of the CDC website, last updated February 12, 2018, listing the four clinical trials assessing PrEP's efficacy: (1) iPrEX, (2) TDF-2, (3) Partner's PrEP, and (4) Bangkok Tenofovir.

*Absent from the screen shot are two additional studies listed as 'PrEP evidence' from aidswatch.org.

CHAPTER IV

LANGUAGE BIAS AND TOPIC MODELING

Though there are numerous documented cases in which numeric data do not reflect the sample from which they are drawn— e.g. the misrepresented data from renowned Harvard scientist Marc Hauser[30] — such cases are almost exclusively discussed within the context of statistical analyses (Aschwanden & King, 2015). Rarely, if ever, is the *language* used to report and supporta problematic numeric data addressed. However, failing to associate problematic language with problematic data unduly ignores an important nuance within problematic science — that language, itself, can be used to misrepresent research data.

Specifically, written language is the medium scientists use to communicate their work. In that communication, contends Gambrill (2012), there are numerous intentionally selected, and unintentionally embedded language patterns that influence how the intended message is interpreted. How the language is contextualized (i.e. rhetorical strategies, surreptitious wording, and withholding of details, among others) directly influences the *quality* and direction of the resulting message. Therefore, if the written language used for reporting scientific work intends to mislead or misdirect an audience, serious implications may result, given the implicit assumptions among the audience that the language used by scientists is both factual and objective.

---

[30] In 2010, Marc Hauser was charged with scientific misconduct due to falsification of data on primate cognition. He would later step down from his professorship at Harvard University, as a consequence of his actions.

Studies on misleading or biased language used to report scientific research is uncommon (Gambrill, 2011). Therefore, how extensive is the use of problematic language in published scientific work is, currently, unknown. However, in a publishing environment in which one science-related article is published every twenty seconds across thousands of academic journals (Bowman, 2014), language bias can be a concern that should not be ignored. Especially, as Sirbu (2015) articulates, it is "writing…[that] plays an important role in the *preservation* of…*realities"* (p.405) making the quality of language much more important.

Therefore, the purpose of this chapter is to explore an understudied component of research bias— *language bias*— by answering the following research question: *Do common bias-inducing factors such as time, funding source, and country of origin, influence latent language patterns in published research*? To answer this question, I will employ technological tools, namely topic modeling, to dissect language via computer algorithms. The overall aims of this chapter are to discuss the prevalence of language bias and answer the research question in a more thorough and objective manner.

Language and Bias

Linell (2012) contends our understanding of language is abstract—that is, language is abstract due to its social construction. Thus, there is no universal way to utilize written or spoken language, as language is bound by different cultural and social norms. Because language is abstract (Evans, 2014), it has a unique power to

communicate anything; ranging from hand-written letters, public speeches, or dissertations.

The degree of power, or influence a given message has when it is communicated is closely tied to *how* that message is presented (Gambril,2011). In other words, in a written text, or in speech, there are numerous rhetorical strategies and linguistic devices one can employ to sway attitudes and opinions on varying subjects. Oftentimes, however, communicators may not be fully aware they are using these strategies in their communication (Andreas 2017).

In all written and spoken language, there are imbedded patterns of word choices. These so-called latent language patterns often have characteristics of misleading or false rhetorical strategies, as they attempt to create a veneer of credibility— especially when the intended message is driven by hidden motives (Abraham, 1995). Gambrill (2011) posits there are two categories of latent language patterns: (1) those that are unintentionally applied in written or spoken language (which she views as unintentional bias), and (2) those that are deliberately placed (what she terms 'propaganda').

The distinction between conscious and unconscious latent language patterns is important as both can pose harmful consequences. Unconscious langue bias, for example, can be emblematic of ignorance (Banks & Ford, 2009). For example, if a researcher holds prejudicial attitudes against a certain minority group, but studies health disparities, there is the potential for manifested language patterns to address certain groups differently, ignore them all together, or provide assessments that do not

accurately reflect that group. Therefore, because of latent prejudicial attitudes that researcher may be, unintentionally, writing about certain groups differently.

Conscious language bias is, more often than not, emblematic of manipulation (Jones, 2016). In conscious language bias, rhetorical strategies, linguistic devices, and withholding of information, among other practices, are deliberately applied to misdirect or mislead an audience (Bricocoli & Cucca, 2016). For example, as it relates to problematic numeric data, several linguistic strategies may be used to mitigate the weaknesses of flawed data. Therefore, conscious language bias offers the potential to *deliberately* promote false or misleading information to attain some sort of gain (be it prestige or financial, among others) (Lakoff, 1995).

One aim of science, however, is to promote a research agenda that is accurate, impartial, and objective, in its relentless pursuit of knowledge (Gordon, 2006). Hubbard (2015) cautions, however, that while such objectivity in science is ideal, not all research agendas are created equally. Some agendas seek to actively mislead, misdirect, or legitimize fraudulent findings as credible science (Ioannidis, 2005).

Only recently, however, have these issues come to light so forcefully (McArdle, 2011). Today, contrasted with the prior history of academia, there are more *documented* cases in which fraudulent science has been identified— evidenced by (1) growing retraction rates (Steen, Casadevall, & Fang, 2013), and (2) a mounting inability to replicate findings (Pashler & Wagenmakers, 2012). In both cases (i.e., replication and retraction) numeric data are, more often than not, identified as *the* primary area of concern (Earp & Trafimow, 2015).

The language employed to communicate those data is often ignored (Egger et al., 1997). According to Linell (2004), however, ignoring how language is contextualized is detrimental to understanding the *scope* of bias in published research. In any academic study, for example, several linguistic devices can be surreptitiously used to alter a message— for supporting fabricated data, for instance. Therefore, by identifying numeric data as the only factor in manifestations of bias – and ignoring language as another important component of bias -- is detrimental to more comprehensive investigations and understanding of research bias (Fairclough, 2013).

As previously mentioned, science should be transparent and objective in all aspects of the research process—reporting included. If instances of fraudulent data passed as credible science are increasing (Marcus & Oransky, 2014), then it stands to reason that language biases may also be prevalent in published research.

Bias-inducing Factors

A bias-inducing factor is anything in the research process that can potentially influence language and create language bias (McArdle, 2011). Though there are certainly numerous bias-inducing variables that can influence language patterns (e.g. see the work of (Delgado-Rodríguez & Llorca, 2004)), I will purposefully explore three of these variables: (1) time, (2) funding source, and (3) nation of origin.

The justification for selecting time, funding source, and nation of origin as bias-inducing factors is as follows. First, I included time because of the ease in which language changes along with research advances (Swaminathan, 2007). Specifically, as new innovations replace older practices, language tends to incorporate and reflect these

changes, over time. Second, I selected funding source as a bias-inducing factor because funding's influence on language patterns is one of the most visible sources of contamination in research (Chopra, 2003; Barden, Derry, McQuay, & Moore, 2006). Finally, I selected nation of origin as a bias-inducing factor for two reasons: (1) there are regulatory differences in research across nations (Van Norman, 2016), and (2) those regulations promote differing opinions and perceptions on medication uptake, which often manifest in language (Arrow & Aronson, 2016).

Current Methods for Detecting Language Bias

As stated previously, in studies of research bias, language is not studied as frequently or as carefully as numeric bias (Gambril, 2011). Two reasons may explain why language bias is understudied: (1) it is not as well-understood as numeric bias, and (2) there are, to date, no systematic approaches to detect biased language.

The primary reason language bias is not as well-understood as numeric bias is the general lack of linguistics training in other social sciences (Linell, 2004). In the social sciences, many fields are primarily data driven and utilize quantitative designs in their research. Therefore, language is not as critically assessed and scrutinized as "hard" numeric data (King, 2011). Linguists, however, are trained to dissect language, and are more attuned to rhetorical devices used to sway language. Therefore, without specific linguistic training, many social scientists are prone to overlooking intentional and unintentional linguistic choices that alter research-reporting language.

Unlike numeric bias— which has objective tools for its detection and measurement (Barden et al., 2006; Chan & Altman, 2005; McArdle, 2011)— for

language bias, currently, no such measures exist. Therefore, identifying something as being biased, with regard to language, remains subjective (Drapeu, 2002). In that subjectivity, if one scholar brings about accusations of bias against another scholar, no matter how credible the evidence, the accused scholar stands on a very stable platform to argue the accuser is biased, him/herself. Therefore, without a more objective and systematic approach to diagnose language bias, accusations of biased work likely amount to little more than contentious arguments amongst scholars seeking to discredit one another's work.

Though an objective approach to detect language bias in research reporting is unavailable in Public Health, Health Education, and Health Promotion, one methodology that could be adapted — popular within the field of Computer Science — is Topic Modeling.

Topic Modeling

What is Topic Modeling? Topic modeling (TM) is a form of text mining via computer algorithms, used to aggregate and segment a large collection of text (known as a corpus[31]) into smaller manageable subsets (Wallach, 2006). The theoretical logic of topic modeling assumes that, in any corpus, there are underlying latent patterns and thematic structures that are difficult to detect due to the sheer volume of information available online and in print (Maldarelli et al., 2014). Topic modeling can be used to

[31] A corpus can be a collection of books, peer-reviewed articles, an aggregate collection of social media posts, or any other form of large-scale textual content. Generally speaking, the larger the corpus of data, the easier to interpret the latent topic.

effectively consolidate such large collections of text and automatically identify only the most important themes embedded within them.

While there are many different forms of topic modeling (Latent Semantic Analysis [LSA], Topic Evolution Model [referred to as CTM], among others), the most popular, widely used, and consistently cited is Latent Dirichlet Allocation or LDA (Blei, Ng, & Jordan, 2003). Unlike other topic modeling algorithms, LDA has been more widely validated in computer science research, and has been used most commonly in topic modeling applications (Hoffman, Bach, & Blei, 2010). Therefore, while other topic modeling algorithms perform nearly the same function, the focus on this section will be on LDA, a math-free discussion of its calculations, and its current applications.

Latent Dirichlet Allocation or LDA. LDA is unique among other types of topic models, because it utilizes Bayesian inferencing as part of its calculus and not basic matrix algebra. As part of Bayes logic, LDA primarily utilizes Gibbs sampling, an iterative inferencing technique that guarantees all data points (in this case words) are represented in the corpus equally (Griffiths, 2002). LDA relies, therefore, on complex probability distributions to compare *each word* (X) *with every other word* (Y) in a vast collection of documents, to determine which words are probabilistically most associated with other words (Wallach, 2006; Steyvers & Griffiths, 2007; Porteous, Newman, Ihler, & Asuncion, 2008).

Words with high probabilities of association are grouped together in a three-dimensional vector space to form latent thematic clusters, while words that provide no

structural meaning – such as prepositions and articles -- are eliminated from the corpus (Wang & Grimson, 2008). The words in each cluster, ideally, are similar enough that one can interpret the thematic meaning of the grouped words with little effort. For example, if we created a corpus on every book written on the topic of pet care, we would likely find several word clusters related to different types of animals. In our hypothetical pet-care corpus, if words such as bark, woof, bone, and fetch were clustered as one latent theme, we would interpret that cluster, or topic, to be about dogs. The same can be said for words such as meow, purr, mice, and feline, being interpreted as a topic about cats.

The logic behind LDA is not unlike the one underlying Exploratory Factor Analysis (EFA), a common technique in exploratory statistics (Péladeau & Davoodi, 2018). EFA, in its purest form, seeks to identify a structure within answers to a survey, for example, based on response patterns from participants (Carroll, 1985). In other words, EFA identifies survey questions answered in a similar enough pattern to group them as a latent factor. Topic modeling, essentially, performs the same task — but rather than use numeric data to create a set of grouped survey items, it relies on patterns among text data to create the latent topics.

Along with being compared to common statistical procedures such as EFA, to help non-computer scientists understand how the algorithm calculates topics, LDA is also equated to a pixelated image (Griffiths & Steyvers, 2004). Below, the readers can see three photos of the Mona Lisa (see Figure 4.1). Each photo, from left to right, is more pixelated than the last. The original unaltered photograph (left) is composed of

millions of individual pixels to provide clarity to the image. The subsequent photos are more pixelated— i.e., pixels have been removed from the original portrait (paradoxically, more pixilation means fewer pixels in the image, overall).

Even with less pixels in the subsequent images, most individuals would feel confident identifying the image as the Mona Lisa. LDA is engaging in an analogous process with a corpus. In other words, the corpus is very large and complex (i.e. the unaltered Mona Lisa). Through the iterative Gibb's sampling process, words are systematically removed until only the most important ones remain. The topics remaining after the LDA procedure is applied, still reflect the original corpus and, more importantly, its content remains identifiable, even with far less information— (i.e. onlookers can still identify the pixelated portrait as the Mona Lisa).

Despite its intuitive similarity with well-established analytic techniques such as EFA, topic modeling is, only now, beginning to be implemented as a valid methodological tool in applied and social sciences (Valdez, Pickett & Goodson, 2018). In other words, though applicable to the social and applied sciences, topic modeling remains, primarily, a methodological tool in Computer Science. Therefore, a brief discussion of *existing* topic modeling applications is enlightening, before examining how topic modeling can be used as a method/tool for capturing language bias in Public Health and health promotion research.

Topic Modeling Applications

Blei and associates (2003) wrote the LDA algorithm to capitalize on, and assign structural meaning to, the explosive growth of online content. Due to its ability

to consolidate large amounts of language with little human input, LDA quickly became noted as the primary platform for text dissection and text mining in the early 2000's (Srivastava & Sahami, 2009).

In its home-field, Computer Science, topic modeling was generally used to explore *how* the Internet was growing and to assess trends in online content, as websites were spawning daily (Perrin & Duggan, 2015). Today, in Computer Science, little has changed in terms of *how* LDA is applied in research. Many recent publications (e.g. Suominen & Toivanen, 2016)) focus, specifically, on social media platforms and advances in LDA's mathematical design.

Nearly two decades after its initial launch, few fields outside Computer Science have adopted topic modeling as a methodological tool. Academic fields that have done so, however, have realized how useful tools such as LDA can be for analyzing large collections of text data. In Business, for example, marketers use topic modeling procedures as a faster and more efficient way to analyze product reviews, such as those on Amazon— something that, at one time, could only be accomplished via focus groups (Titov & McDonald, 2008).

Because TM is not well-known in the social sciences, Valdez, Pickett & Goodson (2018) have called for the adoption of topic modeling as a legitimate methodological tool in social and applied science fields such as Health Promotion, Public Health, and Health Education. More importantly, however, is Valdez and colleagues' contention for not only increased applications of topic modeling in the

social sciences, but also for an increase in the sophistication regarding *how* the method is employed:

> To date, the use of TM [topic modeling] has been limited, and mostly *exploratory* in nature. However, given its ability to unearth the underlying thematic structure of large amounts of data, we contend TM is a powerful tool, applicable to a number of research contexts, especially in the social sciences….While not exhaustive, here we propose three social sciences domains in which researchers could employ and expand the use of TM: (1) as a tool for reducing unintentional reviewer bias in systematic literature reviewing, (2) for practical thematic exploration of qualitative data and thematic analysis validation, and (3) *for comparing similar corpora to explore semantic similarities and differences*. (In-Press, Social Science Quarterly).

According to Valdez and colleagues, topic modeling can be used to answer specific questions by comparing corpora for similarities and differences. These newer applications of topic modeling (e.g. using topic modeling to test specific hypotheses) support the work of Computer Science scholars seeking to innovate topic modeling further, for addressing more complicated and sophisticated research questions (e.g. topic modeling as a form of regression (Wang & Blei, 2011), as well as topic modeling, or pixel modeling, of images (Zeng, et al., 2017), among others).

Although TM is a method to detect nuances of language patterns, its current applications are, mostly, limited to exploratory analyses seeking to identify meaning in a large collection of text. Therefore, the utility, and application of topic modeling in

this investigation will move beyond LDA's intended exploratory roots to answer the following question — *Do common bias-inducing factors such as time, funding source, and country of origin, influence latent language patterns in published research, in ways that can be detected with topic modeling procedures?*

Methods

To answer this research question, I created one grand corpus of data for each bias-inducing factor. Each grand corpus, (1) time, (2) funding source, and (3) nation of origin, was composed of published abstracts from different subjects in Public Health and Health Promotion and Behavior, namely: (1) ADHD medication (to test time), (2) sugar and the human diet (to test funding source), and (3) pediatric/perinatal Highly Active Anti-Retroviral Therapy (P-HAART) (to test nation of origin).

I selected different content for each corpus for several reasons. Chief among them is: each of these subjects, has, at some point, been met with conflicting, divergent public opinion[32]. Therefore, the grand corpora for each bias-inducing variable could easily be divided into competing sub-corpora for subsequent analysis in LDA to detect language shifts, or language bias (Linell, 2004). A second reason I purposefully selected these topics was the volume of publications available in each group. With a more robust corpus, the topics are clearer and more concise than with smaller corpora (Hoffman et al., 2010). Finally, topic modeling, as a method is not biased by topics or content and should be equally useful for any and all subjects.

---

[32] The rationale will be discussed at length in subsequent sections.

Below, I present the composition of each corpus and how each one was treated. Each description provides: (1) a brief justification for its inclusion in this study, (2) the search terms used as inclusion criteria for each corpus, (3) how each corpus was subdivided into a smaller sub-corpus, and (4) the final number of journal article abstracts[33] (the units of analysis, here) in each corpus or sub-corpus.

Corpora

ADHD Medication to Test Time. To observe the effects of time on language, I selected research on the flagship brand name medication used to treat Attention Deficit Hyperactivity Disorder (ADHA): Ritalin. In the early 1970's Ritalin was frequently and widely administered, without a full psychiatric diagnosis, to children exhibiting signs of hyperactivity in classroom and home settings (Morton & Stockton, 2000). Today, due to concerns with prescription drug abuse by teens and over-medicating of children, Ritalin is only prescribed, as a controlled substance, after extensive psychiatric evaluation in severe cases of ADHD (Shillington, Reed, Lange, Clapp, & Henry, 2006).

The specific purpose of the corpus of Ritalin-related research was to determine if research language across time would reflect attitude changes toward pharmacotherapy for ADHD. To compose the corpus, I searched four major databases (1) PubMed, (2) EbscoHost, (3) Web of Science, and (4) Medline. In each database search, I downloaded articles using the following search terms: (1) Ritalin, (2)

---

[33] I elected to analyze abstracts and not entire journal articles because abstracts can be downloaded and stored in MS Excel in a more concise and efficient manner.

Methylphenidate, (3) Concerta, (4) Daytrana, (5) Ritalin LA, and (6) Metadate. These words were used as inclusion criteria because they are the most commonly used brand names associated with Ritalin and Ritalin's generic name, Methylphenidate.

My analysis focused on the text of the research reports' abstracts. The final number of abstracts included for analysis was 5,216 ranging in dates from 1970 to 2018. Those abstracts were, then, further subdivided into decades. In other words, every abstract published between 1970-79, 80-89, 90-99, 2000-09, 10-18, was indexed into a different and unique sub-corpus – one sub-corpus for each decade. Each of these decade sub-corpora was run as a separate topic model to assess (1) how language changed throughout the history of Ritalin research, and (2) if language changes reflected current attitudes regarding Ritalin prescriptions (i.e. over-used and abused study drug).

Funding Source. To observe how sources of funding for research influence latent language patterns, I sought to compare industry- versus federally-funded corpora of studies investigating the role sugar plays in the human diet. The rationale for selecting sugar studies was due to the Sugar Industry's documented history of attempting to sway public attitudes about sugar via questionable and non-replicable research[34] (Aubrey, 2017).

[34] In the 1960's and 70's the Sugar Industry funded a group of researchers at Harvard to downplay the negative health outcomes associated with heavy sugar consumption. The research conducted at Harvard was, then, used as evidence to lobby for dietary recommendations in support of a high carbohydrate, low fat diets (Domonoske, 2016).

More important in my decision to include funding source as a bias-inducing variable, is the academic and medical community's widely supported consensus that chronic sugar consumption leads to poor health outcomes (Fischler, 1987). Therefore, any language differences between industry- and federally-funded studies that advocated for, or mitigated the risk of, sugar consumption could be due to the sugar industry's financial involvement in research.

To compose this corpus, I searched four major databases (1) PubMed, (2) EbscoHost, (3) Web of Science, and (4) Medline. In each database search, I included the following search terms to find articles examining the role of sugar in the human diet: (1) Sugar Diet, (2) Sugar and Diet, (3) Sugar in Diet, and (4) Sugar in Human Diet. The searches yielded a total of 828 abstracts.

Subsequently, I conducted a second and third searches to further eliminate abstracts that did not exactly match the inclusion criteria. Search number two, specifically, sought to find abstracts funded *exclusively* by large-scale, federal national institutes, such as: (1) NIH, (2) CDC, (3) FDA, and (4) NSF— $n = 212$ abstracts. The third search within the original 828 abstracts also sought to find studies, or articles, funded by the sugar industry (common funders included, but not limited to: PepsiCo, Coca Cola, Nestle Inc., Kraft Food Brands, and the National Corn Refiners Association)— $n = 71$. Therefore, the final number of abstracts included for analysis was 283, with the federally-funded studies ($n = 212$) and industry-funded ones ($n=71$).

Nation of Origin. To observe rhetorical, English language differences between nations, I compared abstracts studying P-HAART conducted, funded, published and based either in the United States or in Europe. The reasons for choosing the United States and European nations as moderating variables were: (1) the general credibility of US and European research within the international science community (Grunert & Wills, 2007) and (2) the potentially different attitudes regarding long-term medication interventions for children and adults[35].

To compose this corpus, I searched the same databases as for the previous corpora: (1) PubMed, (2) EbscoHost, (3) Web of Science, and (4) Medline. In each database, I searched the following terms to *only* include investigations studying P-HAART: (1) Pediatric HAART, (2) Perinatal HAART, (3) Paediatric[36] HAART, (4) P-HAART, (5) infant HAART, and (6) Pediatric Highly Active Anti Retro-Viral Therapy. The searches yielded a total of 1,149 abstracts.

Each abstract was further sorted into one of two sub-corpora depending on their nation in which the study was conducted and published in — (1) United States, and (2) European nations. For inclusion in the US sub-corpus, specifically, studies had to have: (1) US authors, exclusively and (2) studied/examined a US-based sample. Europe had

---

[35] In the United States, generally, pharmacotherapy is more trusted and widely applied across a large spectrum of physical ailments. The attitudes in the US stand in contrast to those of European nations who often view pharmacotherapy as a last option, especially with regard to children (Marazzi et al., 2006).

[36] Paediatric HAART is the correct spelling for 'Pediatric" in the United Kingdom.

the same criteria: for inclusion in the European Union sub-corpus, studies had to (1) be written by authors from the EU, exclusively, and (2) have obtained their sample(s) from EU nations.

However, few of the indexed abstracts conducted investigations on individuals in *either* the United States or in Europe— the majority of the studies were executed either by an international team, or among populations *outside* of the US or Europe (Rosen & Fox, 2011). If this was the case, or if there was any ambiguity where the studies originated, they were eliminated from the analysis. Therefore, the final corpus was significantly smaller than the preceding two corpora (assessing time and funding source): United States, 74 abstracts, and 56 European abstracts. Fortunately, according to Blei (2003), the overall word count in each sub-corpus was enough to generate an interpretable topic model.

Analyses

I ran a separate topic model for each sub-corpus addressing the three bias-inducing variables. For time, a total of five topic models were generated (one for each decade: 1970-79, 80-89…2010-2018). Funding source required two topic models (industry- and federal-funding) and nation of origin also required two topic models for comparative analysis (United States-based studies and European-based studies).

All analyses were conducted using statistical programming software R version 3.4.2 and the following downloadable R packages: (1) topicmodels (*sic*), and (2) tm (*sic)*. Both packages —which are specialized program extensions for R—allow researchers to run text data through a multi-step process to prepare for analysis. Such

preparation includes: (1) removing punctuation, numbers, special symbols (e.g. *, <, >, &, among others), (2) stemming the document (i.e. removing all suffixes from words so that only the root word remains), and (3) creating a document term matrix (dtm), which is an aggregate calculation of how many times every word is used in a corpus, or sub-corpus.

For ease of interpretation, clarity, and consistency, I generated a similar 5x10 structure topic model for every sub-corpus. In other words, each sub-corpus, regardless of the actual final word count, was stripped down to only include *the five most important topics* with the *top ten associated words* in each topic. The decision to utilize a 5x10 matrix structure is supported by the theoretical positioning of Blei and colleagues (2003) who contend that, due to the lack of 'fit' statistics in topic modeling methodologies, researchers have full right to generate any number of topics per topic model deemed necessary to accurately interpret its structure.

To account for a lack of fit statistics, I opted to, instead, assess inter-rater reliability of the results— which is a check for overall consistency among scholars in interpretations of qualitative data, such as interviews (Armstrong, Gosling, Weinman, & Marteau, 1997). I presented the topic models, without prior discussion, to a qualitative methodology researcher. Upon finding a mutually agreeable interpretation of the context of each topic model, extraneous words deemed unnecessary to the overall dialogue were removed from a final analysis. The final models were transposed into Excel, and saved as one file along with the original text file for each respective sub-corpus.

Results

Time

The purpose of this analysis was to assess *how* language employed in the reporting of Ritalin studies changed over time, and to speculate on the reasons fueling the changes. For ease of detection, I divided the corpus into decade-spanning sub-corpora to compare differences among decades (see Table 4.1).

Across each decade, the most important theme (i.e. all topics listed in bold) was consistent. In other words, probabilistically speaking, in each decade, each sub-corpus was primarily over the same content— an obvious finding considering the corpus was built from Ritalin research. More important was the general stability of the content over time.

Specifically, each of the most important topics (i.e. the topics in bold) across all decades contained a fairly similar amount of words in their respective models: *methyl, disord, adhd, mph, effect, behavior, drug*, among others. Other words in the most-important topics contained similar clinical-type language, narrowing in on Ritalin's effects on behavior, dosing, stimulants, other cognitive functions. Therefore, the general composition of the corpus reflects, primarily, concerns with testing and Ritalin effects.

The remaining topics throughout the span of Ritalin research (i.e. other topics that are *not* the topic in bold), changed gradually. In the 1970's and 1980's *children* and *boy* were the populations most frequently addressed in the studies; frequent enough

to emerge as part of a thematic latent topic. Beginning in the 1990's however, other populations began to emerge often enough to appear within other latent topic (e.g. parent, human, and rodent) and, in the 2010-18 sub-corpus, the term *adult* appears. Other words, such as *abuse, toxic, addict*, begin appearing in the 1990's and were absent in older sub-corpora topic models (see Table 4.2). Below, Table 4.2 from the document term matrix, depicts the rankings of words by importance (i.e. how often words are used in a sub-corpus).

At the beginning of the Ritalin era, for example, the 79th most important word, *boy*, reflected the only population being tested— *girl, adolescent* and *adult* did not appear in that decade's sub-corpus at all; during that time few studies focused on groups other than boys. Subsequent decades saw diversification regarding *who* was tested, eventually including girls, adults, and adolescents. Beginning in the late 90's, and extending into the 2010-18 decade, the terms *adult* and *adolescent* became *more* important (i.e., more frequent) than the original 1970's term '*boy*.' Further, words such as *abuse, adverse, and side* (as in "side- effect") also gained importance and became much more visible over time.

Funding Source

This analysis sought to determine if the source of funding (i.e. industry versus federal funding) would have an effect on latent topics when studies reported testing the same hypothesis— *Increased sugar intake is not associated with negative health outcomes*. Overall, the language in both topic models was notably different (see Table 4.3.).

In the topic model for federally-funded reports, the most important topic (in bold) contains the words *diet, food, sugar, intake, increase, weight, high consumption, energy,* and *risk*. Further, the other topics in that model, particularly Topics Two, Four, and Five, contain clinical-type language associated with outcomes related to sugar consumption (e.g. *metabolism, disease, insulin, effect, mice, liver, link, bod, tumor*, among others). Topic Three was notably different from the other topics. Rather than language reflecting clinical outcomes, Topic Three centered on interventions and cost (e.g. *program, nutrient, polici, cost, ssb* (a frequently used acronym for sugar sweetened beverages), *ses* (socio-economic status), and *regress*.

The industry-funded topic model, had a different emphasis altogether. The computer-identified most important topic, Topic Five, contained the following words: *intake, sugar, diet, cosum, food, energy, beverage, consumpt, dietary*, and *pattern.* Diet, as in food consumed daily, was a recurrent theme in the majority of the remaining topics, especially observable in topics Two, Three and Four. In those topics, food related words such as, *calori, effect, baseline, promote, breakfast, fruit, juice, eat*, among others, were also common.

Topic one in the industry-funded topic model, was notably different from topics Two, Three, Four, and Five. Rather than emphasize diet— as in food consumption[37]— Topic One uniquely discussed outcomes of sugar consumption such as increased

[37] A cursory read-through of the articles composing the industry-funded topic model will show the majority of the articles define diet as 'daily food consumption.'

adiposity and heart function (e.g. *total, increase, fructose, reduce, eat, obes, cvd* (cardiovascular disease), among others). Therefore, though the most recurring theme in the topic models was diet, there was still some mild variation in published content in the industry funded topics, as well. However, the industry model was generally less specific, overall with regard to physical outcomes of sugar consumption, than the federally funded model.

Nation of Origin

This analysis sought to determine if P-HAART studies conducted, funded, and published in the United States and other European nations would produce differing topic models. As with the previous analyses, the language for both models differed (see Table 4.4).

Language from the US-based studies centered on two foci: (1) the prescribing and administering of P-HAART to infants, and (2) general recommendations for infants potentially exposed to HIV. The computer-identified most important topic for US-based studies contained the following words: *HIV, infect, children, pediatr, health, report, care, youth, infant*, and *disease*. The remaining topics (e.g., topics Two through Five) were similar to Topic One with regard to population and scope in their respective topics. In other words, much of the content in the corpus supporting this topic model was primarily centered on uptake.

Specifically, Topics Two and Three in the US-based model focused on HIV transmission and pharmacotherapy applications, using words such as: *drug, medic, birth, issue, viral, test, aid, born, recommend, high, dose, exposure, matern*, among

others. Topics Four and Five used slightly different words to convey a focus on general recommendations and federal guidelines, such as: *regimin, factor, cdc, human, research evalu, adult*, and *patient*. Much of the language in the US-based topic model also reflects the national recommendations for PHAART in the United States:

> The uses of ARV regimens in newborns include:
>
> • ARV Prophylaxis: The administration of one or more ARV drugs to a newborn without confirmed HIV infection to reduce the risk of HIV acquisition.
>
> • Empiric HIV Therapy: The administration of a three-drug combination ARV regimen to newborns at highest risk of HIV acquisition. Empiric HIV therapy is intended to be preliminary treatment for a newborn who is later confirmed to have HIV but also serves as prophylaxis against HIV acquisition for those newborns who are exposed to HIV in utero, during the birthing process or during breastfeeding and who do not acquire HIV.
>
> • HIV Therapy: The administration of a three-drug combination ARV regimen at treatment with HIV diagnosis" (National Institutes of Health, 2017, pg. H-1, retrieved at https://aidsinfo.nih.gov/contentfiles/lvguidelines/PediatricGuidelines.pdf)

The European studies also had two foci: (1) management and diagnosis of HIV, and (2) guidelines and recommendations for P-HAART. The computer-identified most

important topic contained the following words: *art, parent, manage, status, diagnos, drug, screen, europ, hundred.* The remaining topics (e.g. Topics One, Three, Four and Five) were notably different from one another, with little overlap in content.

Topic Three, for example, discusses guidelines using the words *guideline, health, recommend, provid, migrant, aid, adolescent,* and *disease.* Topic One, on the other hand, as addresses national reports of HIV infection: *year, present, patient, country,* and, *report*. Topic Five can further be interpreted as care for children living with HIV: *HIV, children, infect, paediatri, care, age, women, clinic, follow, European.*

Overall, much of the language in the European topic model also reflected European recommendations for P-HAART from the Paediatric European Network for Treatment of AIDS (PENTA) revised 2015 guidelines:

> PENTA guidelines seek to optimize treatment for children in Europe. However, particularly during adolescence, care may need to be individualized. This document should not be seen as a standard for litigation as individualization of case management and departure from this guidance may be necessary and indicated. Significant changes since the 2009 guidelines include:
>
> - decreased frequency of laboratory monitoring in clinically stable children both on and off ART;
> - consideration of ART initiation in all children aged 1–3 years in order to minimize the risks of disease progression or death;
> - consideration of ART initiation at higher CD4 thresholds in children > 5 years of age in order to optimize potential for immune reconstitution;

- additional clinical indications for ART initiation at all ages;
- addition of newer protease and integrase inhibitors to first-line preferred and alternative third agent options, respectively;
- update on specific guidance in the context of hepatitis B and C virus and tuberculosis (TB) coinfection in light of new ART options at younger ages;
- a summary of new drugs [including new fixed dose combinations (FDCs)] that can be considered for second- and third-line options and of the 'pipeline' of new drugs likely to become available;
- an emphasis on the needs of older children and adolescents as they approach transition to adult care (Bamford, et al., 2015, p. e5)

Discussion

The purpose of this chapter was to explore an understudied component of research bias, namely, *language bias*, by answering the following research question—*Do common bias-inducing factors such as time, funding source, and country of origin, influence latent language patterns in published research, in ways that can be detected with topic modeling procedures?*

To varying extents, the language examined for each bias-inducing variable changed among rival sub-corpora, as highlighted above. However, in order to appreciate the nuances in the language differences among the topic models, one must also understand the potential underlying factors. The discussion to follow will highlight, among other points of consideration, what factors, circumstances, or events

may have influenced or swayed language patterns to create the documented differences among the topic models.

Time

One source of language changes in Ritalin research may be the growth in research publications, and evolution of research trajectories, on Ritalin and Methylphenidate over time (see Figure 4.2). In the 1970's there were less than one hundred scientific publications on Ritalin, the majority of which were *clinically* focused (e.g. the most important topic for that decade contained the words *methylphenidate, dose, patient, effect, interv*, among others). In the 2010 decade, the number of publications increased to nearly 3,000 clinical, psychological, sociological, and epidemiological research studies on Ritalin and other ADHD pharmacotherapies. Thus, there were inherently *more* discoverable latent topics in more recent years of Ritalin research.

The prolific growth/evolution of methylphenidate research, itself, along with subsequent language changes, can be explained through varying historical contexts. Chief among them is the changing demographics of the first cohorts administered Ritalin to treat hyperactivity. Specifically, the first children to whom Ritalin was administered in the early 1980's, grew into adulthood in the 1990's and early 2000's (Schacter, Pham, King, Langford & Moher, 2001). While this drug was intended primarily to treat children (specifically, *boys*), the shift in demographics prompted renewed testing to determine if Ritalin regimens should continue into adulthood (Cox, Merkel, Kovatchev & Seward, 2000). The desire to increase the scope of Ritalin was

equally reflected in the topic models— in later years, the terms girl, adolescent, and adult eventually emerge as salient enough to be captured by the models.

Due to successful efficacy and safety testing among adolescent and adult populations, guidelines governing Ritalin and other ADHD pharmacotherapy adapted to include a patient population that did not consist merely of children (Gamble, 2011). For example, in 2001, guidelines published in the Journal of Pediatrics noted the appropriate age for Ritalin use *was no younger than* six years of age and *no older than* twelve. In an update to those guidelines (in 2011) there were two major changes: (1) ADHD was now reclassified from a psychological disorder into a chronic condition and (2) the appropriate ages to administer Ritalin were changed to include children as young as *four* and adults *eighteen and over (Journal of Pediatrics,* 2011).

With adults, adolescents, and children now using Ritalin, the amount of prescriptions written for ADHD pharmacotherapy doubled within one decade (Boffey, 2015). And, due to the wide availability and administrations of Ritalin and other ADHD pharmacotherapies, researchers were further able to document new aspects of Ritalin that were previously unstudied— such as its negative outcomes. Specifically, in the 1990's Ritalin—once considered a safe drug intended to treat hyperactivity in children— was now classified as a high-risk study drug linked to abuse (Babcock & Byrne, 2000; Morton & Stockton, 2000). More importantly, the topic model was able to capture this important nuance— the term abuse would first appear as a topic in the 1990's model.

The topic models in every decade, therefore, were able to accurately capture changes in perceptions, as well as current events within Ritalin research. For example, the changes in scope (e.g. changing populations from children only, to children, teens, and adults), the link to its abuse, and concerns over safety. Though the evolution of Ritalin research, itself, is not indicative of bias within the field, the important take-away is that technological tools, such as topic modeling, correctly assisted in creating latent topics reflecting important events *at that time, and within a given decade.* TM's ability to detect these changes can be useful for historically assessing a given field's bias(es) regarding overall perspectives, targeted population groups, treatment preferences, and other important patterns governing research and practice at different times.

Funding Source

The topic model for federally-funded research contained more medically-styled language that was more critical of sugar than its industry-funded counterpart Specifically, in the model for the federally-funded studies, words such as *risk, weight, gain, tumor, insulin, metabol,* and *disease* can be interpreted as emblematic of consensus within the health care community, regarding sugar— excessive consumption has negative health outcomes (Rodearmel et al., 2007). Further, similar medical-style language is also seen in federal guidelines, such as those from the CDC, to call for decreased sugar consumption by children and adults (Park et al., 2014).

This clinical language, however, is absent altogether in the topic model for the industry-funded research. Instead, language in that topic model appears to treat sugar

merely as another variable. In this model, sugar is treated as a normal part of the human diet, to be enjoyed in moderation. More importantly, language in each of the topics in the model for industry-funded research tended to pair sugar with other household items and behaviors often billed as healthy— such as *fruit, juice, grain,* and *breakfast*.

Gambrill (2011), who contends industry-based investigations are inherently biased, calls diverting attention away from serious outcomes, "oversimplification [used to] dull critical thinking" and mask lingering controversies (p.289). Wolfson (2017) further adds that oversimplification is common among *many* types of research funded in-house, in a bid to mitigate a bad reputation. For example, in 1993 Congress was set to pass legislation banning public smoking across all public establishments. Major tobacco manufacturers, concerned over potential profit loss, funded their own investigations intending to downplay the role of second hand smoke for adults and children. As with the sugar industry's diverting away from outcomes and, instead, re-directing sugar to associations with health foods, tobacco manufacturers sought to divert attention away from health consequences of second-hand smoke and contended smoke-free establishments will only lead to decreased profit margins for public establishments, such as bars and restaurants. Thus, it is apparent the sugar industry, similar to the tobacco industry, is attempting to misdirect the audience by shifting the narrative away from health implications of sugar consumption, to focus, instead, on healthy foods and healthy behaviors, such as eating breakfast.

Also absent from the model for industry-funded research, but evident in the federally funded studies, are the words *SES, cost,* and *ssb*. Though those words could

be interpreted in various ways, their absence in the model for federally-funded studies may indicate the industry's lack of concern for general societal-level implications of sugar consumption. Park et al., (2014) contend sugar consumption is highest among lower-income, minority families due to its affordability. In other words, because processed foods and sugar-sweetened beverages are often more affordable than natural, less-sugary products, families of lesser means are more likely to purchase these goods. Consequently, contends Pechey (2016), these families are likely to suffer *more* co-morbidities and sugar-related chronic conditions than individuals who consume less sugar. Therefore, the words *obes, metabol, liver, disease*, appearing alongside, *ssb, SES*, and *cost* in the federally-funded model indicates federally-funded research may have focused more on the psychosocial and economic aspects of sugar than their industry-funded counterparts.

When examining the remaining topics in the topic model for federally-funded research, one can infer that discussing SES and cost is a small part of the larger narrative— with its nearly-exclusively focus on negative effects of sugar in the human diet. In the industry-funded model there is an absence of vocabulary describing co-morbidities and important information on cost, alongside the presence of language pairing sugar with healthy products. Topic modeling was able, therefore, to capture these bends in the research agendas and their manifestations in latent writing patterns.

Nation of Origin

The US topic model for Nation of Origin was clear regarding the targeted population: infants and children (e.g. *birth, issue, virus, transmiss, infant, hiv*,

children). More evident was the sense of urgency in administering P-HAART at the time of birth: *birth, issue antiretrovir, prophylaxi, receive.* In the European TM, however, the target population was not as clear, as there were more emergent groups throughout the corpus and final topic model: *provid, migrant, aid, adolesc, women, children.* Absent altogether from the European model was the word *infant* despite this corpus being composed of studies regarding *pediatric* HAART.

As mentioned previously, these distinctions most likely stem from the regulatory differences between the United States and other nations in the EU. Some of those differences, discussed below, lead to conflicting attitudes regarding medication uptake, and shape the type of research being done. In the US, medication uptake is viewed favorably, as evidenced by the fact 70% of Americans report taking at least one prescription drug daily (Mayo Clinic, 2018). In most nations in the EU, less than 40% of their populations report taking at least one prescription drug, daily—suggesting, perhaps, a more cautious and skeptical stance toward medications (Eurostat, 2014, retrieved at: http://ec.europa.eu/eurostat/statisticsexplained/index.php/Medicine_use_statistics#Prescribed_medicines.)

EU pharmaceutical research and medication distribution are regulated heavily by the government. This is done, in part, to de-incentivize profiteering by pharmaceutical companies (Eger & Mahlich, 2014). For example, when compared to the United States, almost all major medications are significantly cheaper in the EU (Danzon & Chao, 2000). More importantly, less research and development of

pharmaceuticals occurs in Europe due to smaller profit margins when compared to the US (Filson & Masia, 2007). Therefore, due to regulations in which profit incentives are removed, any tested medication in Europe will be more widely scrutinized, evaluated, and thoroughly tested before ever being approved for use among the general population (Eger & Mahlich, 2014).

The United States' federal regulations allow for more competition between governmental outlets, such as the NIH and CDC, and industry-based counterparts—which infers more of an open-market structure (Spiegel, 1991). In fact, over the last several decades, US-based companies outspent the federal government's budget by over 1 billion USD (Pharmaceutical Manufacturer's Association Annual Report (1989; 2017). And, because industry is outspending the federal government, many governmental agencies such as the FDA and CDC will often defer to findings from industry-funded clinical trials (Carr, 2017). Such actions result in medications that are heavily marketed, more expensive, and difficult to obtain without private insurance (Lyles, 2014).

Returning to the topic models I presented, those for the US seem to reflect the same sense of urgency characterizing the National Institutes of Health's guidelines for P-HAART, which recommend "Empiric HIV therapy is intended to be preliminary treatment…without consent to newborns who are exposed to HIV in utero, during the birthing process or during breastfeeding and who do not acquire HIV" (NIH, 2017, H-5) . Such urgency may well stem from the prevailing view to "hit HIV hard and early"

and prevent HIV transmission as early as possible (Ho, 1995, p.450) but may also be shaped by the market-oriented culture of drug policies in the US.

On the other hand, the European model also reflected European national guidelines which, unlike the US guidelines' focus on newborns, suggest, "consideration of ART initiation in all children aged 1–3 years in order to minimize the risks of disease progression or death …and considerations of older children into adulthood" (Bamford et al., 2015, p.e5) . The different emphasis on who should receive care most likely stems from the more regulated and scrutinized style of EU-driven research, in tandem with the culture of skepticism toward over-use (and, most likely, early-uptake) of medications. In this case, as with time and funding source, the topic models were able to capture those nuances, peculiar to each geographic region and culture(s).

Conclusion

Though authors such as Monroe (2013) contend embedded messages and surreptitious wording *are* acceptable in the scientific literature, continued reliance on such practices have the potential to lead to further erosion in the credibility of science. As Hubbard (2015) cautions, while all scientists have an agenda, not all agendas are created equally, and certain agendas seek profit over progress. Therefore, in a world full of problematic biased language, we should be better equipped to address and mitigate biased language from science-based literature, if and when it is present. One methodology presented in this chapter, topic modeling, is a tool that can assist scholars in dissecting and interpreting language in ways that help identify biased patterns.

This chapter sought to elucidate the importance of an overlooked component of bias— language. To further cement language bias as a legitimate component of research bias (see Chapter 2 for further development), I briefly outlined examples of three bias-inducing factors: (1) time, (2) funding source, and (3) nation of origin. After outlining the importance of each factor, I tested topic modeling, a method common in Computer Science, for detecting language biases associated with these (and other) bias-inducing factors.

I used topic modeling to accomplish two objectives: (1) frame topic modeling as a legitimate methodological option in Public Health and all fields conducting health-related research, and (2) to answer an important question— *Do common bias-inducing factors such as time, funding source, and country of origin, influence latent language patterns in published research, in ways that can be detected with topic modeling procedures?*

After the analyses of various topic models for each of the factors, the answer is *yes*, those factors do influence language patterns. Yet, most important was how I arrived at the answer— by using technology to parse out only the most important information from a large collection of research. The most salient finding from the analyses presented here is: not only was topic modeling able to identify differences, but these differences also indicated time-related trends, policy recommendations, and rhetorical strategies used to highlight or down-play researchers/scholars' preferences. In other words, the findings from each topic model demonstrated how scientific language aligns itself with the larger narratives in which it is embedded. These findings

suggests topic modeling is successful at consolidating a large amount of information into manageable chunks, and in capturing nuances of meanings, as well as preferred patterns of language used to communicate this information.

Therefore, at the very least, findings from this dissertation chapter support the notion that topic modeling should be studied further as a valid tool with which to detect language bias and, ultimately, identify potentially harmful research biases. Though still unique for Public Health, tools such as topic modeling have assisted Computer Science scholars with assigning meaning to abstract "noises", such as online content. Thus, we should continue borrowing this and other potentially useful tools to further advance studies of bias— both numeric *and* language based.

Though language bias is certainly a newer (and less studied) component of research bias, its novelty does not imply lack of importance. As Gambrill (2011) contends— the key to making an informed choice is access to quality information. She further adds, many consumers are unable to distinguish varying levels of information quality, due to the saturation of 'propaganda,' or misleading messages. Because the lay public cannot differentiate between good and bad quality science, that obligation-- contends Ioannidis (2005)— falls upon scientists who must uphold the credibility of their endeavors. Fortunately, with tools such as topic modeling, studies on language bias are now much more accessible, and if ever so slightly, a little less subjective.

Figure 4.1. A progressively pixelated Mona Lisa is an analogy showcasing how topic modeling takes a large collection of text content, simplifies it to only the most important parts, but still maintains the overall structure of the original text data.

Table 4.1. Topic Models on Methylphenidate Research by Decade

	1970-79				
	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
1	children	**methyl**	attent	perform	group
2	hyperact	**effect**	behavior	learn	rate
3	drug	**behavior**	medic	normal	motor
4	hyperkinet	**find**	measur	ritalin	found
5	improv	**stimul**	condit	treat	height
6	arous	**abstract**	control	differ	subject
7	respons	**hyperact**	problem	compar	affect
8	report	**physiolog**	test	neurolog	present
9	treatment	**show**	dose	task	case
10	weight	**respond**	age	dextroamp	cognit

	1980-1989				
	Topic 1	**Topic 2**	Topic 3	Topic 4	Topic 5
1	studi	**methyl**	measur	drug	attent
2	task	**hyperact**	behavior	disord	differ
3	test	**children**	respons	treatment	rate
4	condit	**effect**	hour	activ	stimul
5	assess	**boy**	cognit	concentr	pharmacolog
6	ritalin	**deficit**	improv	subject	present
7	administr	**perform**	process	medic	meal
8	mgkg	**dose**	prolactin	add	time
9	reaction	**increas**	interact	growth	effect
10	group	**studi**	control	posit	design

	1990-1999				
	Topic 1	Topic 2	Topic 3	Topic 4	**Topic 5**
1	parent	diagnosi	function	effect	**methyl**
2	abus	academ	three	children	**disord**
3	assess	diagnos	human	drug	**adhd**
4	amplitud	remain	stimulus	hyperact	**attent**
5	experiment	latenc	sensit	deficit	**behavior**
6	potenti	addit	consist	ritalin	**stimul**
7	appear	edsub	stimuli	medic	**respons**
8	attribut	emiss	therapeut	patient	**dose**
9	comparison	issu	tomogra	test	**mgkg**
10	deficit	lower	addict	cocain	**report**

	2000-2009				
	Topic 1	**Topic 2**	Topic 3	Topic 4	Topic 5
1	focus	**methyl**	mode	attent	fluoxetin
2	pfc	**adhd**	extens	patient	conflict
3	characterist	**effect**	afternoon	dose	secondari
4	neuropsycholog	**mph**	error	hyperact	communic
5	distribut	**children**	noradrenerg	improv	neurotransmiss
6	place	**disord**	randomis	differ	rodent
7	biolog	**drug**	sexual	year	selfreport
8	toxic	**increas**	pathway	assess	bid
9	locat	**medic**	therebi	suggest	blind
10	valu	**stimul**	antidepress	includ	continu

Table 4.1 Continued Topic Models on Methylphenidate Research by Decade

	2010-2018				
	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
1	**adhd**	improv	roi	neurochem	adult
2	**methyl**	day	aetiolog	basic	perform
3	**mph**	psycho	fertil	genotox	present
4	**effect**	case	produc	belief	function
5	**disord**	cocain	snps	site	baselin
6	**patient**	receptor	subcort	cage	task
7	**drug**	male	cmethyl	dawley	administr
8	**children**	common	fix	frontostriat	atomoxetin
9	**medic**	investig	fli	abl	particip
10	**increas**	time	pkc	arrest	receiv

Table 4.2. Word Ranking by Decade on Methylphenidate Research:

	1970-1979	1980-1989	1990-1999	2000-2009	2010-2018
Boy	79	14	37	173	233
Girl	-	581	373	426	606
Adult	-	275	71	32	21
Adolescent	-	430	97	51	33
Toxic	673	-	551	704	931
Side	220	-	26	169	169
Adverse	179	-	231	97	129
Abuse	-	-	-	3219	104

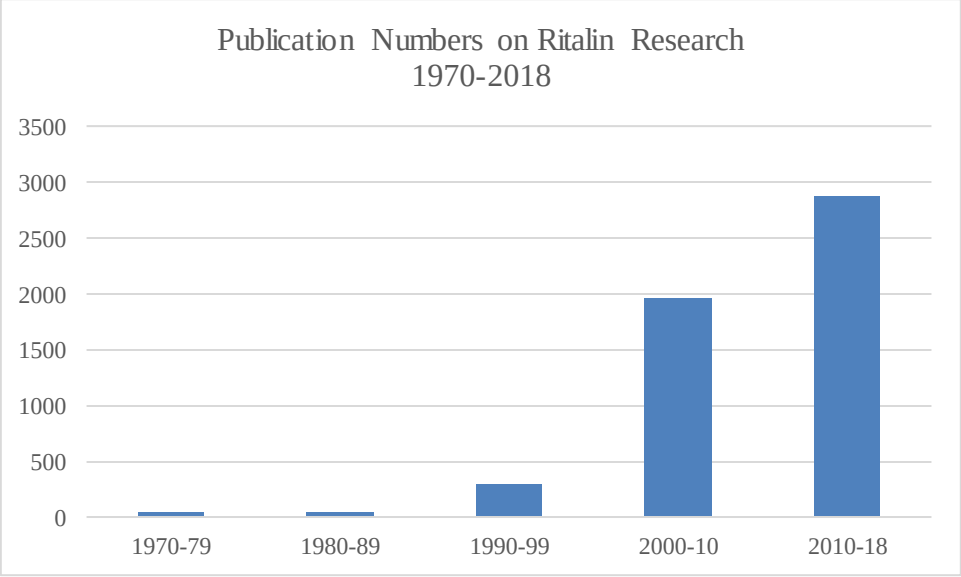
Table 4.3. Topic Models for Industry- and Federally-Funded research reports on sugar's role in the human diet

	Industry						Federal				
	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5		Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
1	total	fruit	chang	calori	intak	1	diet	fructos	genet	resist	famili
2	increas	weight	product	effect	sugar	2	food	beverag	program	signal	individu
3	fructos	mean	reduct	calor	diet	3	sugar	metabol	nutrient	term	random
4	reduc	breakfast	design	lower	consum	4	intak	diseas	lower	larg	women
5	eat	women	effect	trial	food	5	increas	obes	polici	home	store
6	obes	blood	baselin	free	energi	6	weight	mice	regress	link	amount
7	well	contribut	promot	examin	beverag	7	high	effect	cost	bodi	insulin
8	cvd	school	measur	obes	consumpt	8	consump	insulin	ssb	gain	loss
9	loss	carbohydr	breakfast	respect	dietari	9	energi	relat	ses	progress	analyz
10	grain	juic	either	observ	pattern	#	risk	liver	analys	tumor	healthi

Table 4.4. Topic models for research on P-HAART in Europe and the United States

	Europe					United States					
	Topic 1	**Topic 2**	Topic 3	Topic 4	Topic 5		**Topic 1**	Topic 2	Topic 3	Topic 4	Topic 5
1	year	**art**	therapi	guidelin	hiv	1	**hiv**	drug	test	antiretrovir	prevent
2	present	**parent**	unit	health	children	2	**infect**	medic	aid	prophylaxi	research
3	patient	**manag**	survey	recommend	infect	3	**children**	birth	born	expos	compar
4	countri	**status**	pediatr	provid	paediatr	4	**pediatr**	issu	differ	receiv	evalu
5	report	**diagnos**	count	migrant	care	5	**health**	virus	recommend	increas	adult
6	start	**drug**	develop	aid	age	6	**report**	famili	high	guidelin	particip
7	antenat	**escmid**	four	adolesc	women	7	**care**	mhps	behavior	regimen	physician
8	acquir	**screen**	time	diseas	clinic	8	**youth**	transmiss	caregiv	factor	present
9	activ	**europ**	case	live	follow	9	**infant**	viral	exposur	cdc	patient
10	differ	**hundr**	childhood	mortal	european	10	**diseas**	resist	matern	human	assess

Figure 4.2. Number of published studies on Ritalin by decade.



CHAPTER V

CONCLUSION

This dissertation sought to raise awareness of the complexity of bias in today's research climate. More importantly, this dissertation also aimed to introduce, and subsequently test, novel technological tools for bias detection in published health-focused research. To accomplish these objectives, I utilized a journal-article formatted dissertation encompassing the following content: Chapter 1— introduced the dissertation, its format, and each chapter's general content area; Chapter 2— called for an expansion of scope of research ethics, to include bias—both numeric and language-related--as a component as worthy of attention as the ethics surrounding human and animal subjects participating in research; Chapter 3— demonstrated the utility and effectiveness of GRADEpro as a tool for detecting numeric bias in published reports, and Chapter 4—demonstrated Topic Modeling as a language-bias detection tool. Below is a brief conclusion of each study followed by practical recommendations for future researchers.

Bias as a Component of Research Ethics

Chapter 2 sought to argue that bias, defined as any factor that influences the quality of research in both published and unpublished work, is a contemporary ethical concern in today's research climate—as vital for the credibility of the scientific enterprise as concerns with the rights of humans and animals participating in research. A secondary goal of this study was to spotlight two types of bias— namely, numeric

and language bias-- and examine how these biases manifest themselves and influence research quality.

After a brief discussion of current problematic issues in science— such as the replicability crisis, among others— I sought to argue bias can potentially manifest within numeric bias and language bias. More importantly, I further argued how current measures intended to address such biases are insufficient, because bias detection, itself, can be inherently subjective.

Precisely because bias detection is subjective, and the overall goal of this dissertation is to promote the discussion around bias mitigation, I introduced two tools that could more objectively assist in the detection of bias— GRADEpro, for detecting numeric bias, and Topic Modeling, for language bias. Chapter 2 concluded with a call to better understand bias and to further test GRADEpro and Topic Modeling as "bias detectors". This call further served as the launching point for the studies presented in Chapters 3 and 4.

Numeric Bias Identification with GRADEpro

The purpose of Chapter 3 was to elaborate on numeric bias— or the bias that can be embedded in the reporting of numeric data. This chapter also sought to test GRADEpro, and the Cochrane Evaluation criteria, as a potential tool for detecting numeric bias. To test and exemplify the use of GRADEpro, I assessed a total of four, large-scale, randomized control trials. These studies tested the effectiveness of PrEP to prevent HIV transmission among non-infected people at high risk for infection. I employed the GRADEpro criteria and software to (1) determine the overall quality of

the evidence presented in the studies' reporting, and (2) to develop a summary of findings table. In tandem, these two steps provide a more objective "picture" of the studies' reporting quality, and constitute the main features of the GRADEpro assessment.

The four studies varied in quality. Studies with lower quality were identified as having one, or more, infractions in one or more of the five evaluation domains: (1) Risk of Bias, (2), Imprecision, (3) Indirectness, (4) Inconsistency, and (5) Publication bias. The summary of findings table served to further interpret data presented in the four studies. Findings from this analysis indicated the efficacy of PrEP varied and, because only p-values were reported when comparing treatment and control groups on side-effects, the studies lacked important information (such as CIs and ES) on the clinical significance of the differences detected (all group differences regarding side-effects were reported as not statistically significant).

Though not entirely objective— much of the assessment of quality involves making value judgements based on a set of criteria, GRADEpro still proved useful for analyzing large clinical trials, systematically. More importantly, the analysis engendered by using GRADEpro elicited important findings within the four analyzed trials that would not have otherwise been visible. Therefore, as it relates to numeric bias and problematic reporting of numeric data, GRADEpro is an appropriate tool and should be utilized in health-focused research in both the social and medical sciences.

Topic Modeling as a Language Bias Detection Tool

The purpose of Chapter 4 was to test Topic Modeling as a potential tool for detecting language bias— defined as the observable and non-observable language structures that shape a message with the intent of misleading the message recipients. To determine if Topic Modeling was an appropriate tool to detect language bias more objectively, I selected three factors commonly associated with language bias— (1) time, (2) funding source, and (3) nation of origin—to test the tool's performance.

I further selected three research topics (Ritalin, sugar and the human diet, and the Pediatric Highly Active Anti-Retroviral Therapy [P-HART]) with which to test changes in language when parsed out among the bias inducing factors. In other words, I examined: (1) changes in language used in Ritalin research over time, (2) differences in how language is used when reports regarding the role of sugar in the human diet are funded by industry or federal sources, and (3) differences in language used by US and European researchers to report on and recommend P-HAART.

Within each of the bias inducing factors, language changes were observable. With Ritalin, there was clear change in language, over time, as more populations are prescribed Ritalin. More importantly, vocabulary changes also attested to the drug's association, currently, with abuse and long-term neurological effects. Regarding funding source, studies funded by the sugar industry employed language that downplayed the harms of long-term sugar consumption. Federally funded studies, however, were more likely to associate sugar with long-term chronic illnesses. Regarding P-HAART, US-based studies were more likely to recommend onset at birth

and without consent. Language in European studies emphasized, instead, older age groups and different populations (e.g., adolescents and immigrants).

Overall, Chapter 4 had two main findings: (1) language bias *can* be objectively detected, and (2) topic modeling, a set of computer-based algorithm, can assist in bias mitigation when large collections of text are parsed out into different groups and compared to one another. Therefore, regarding studies of language bias, topic modeling can and should be implemented as a valid method in the health sciences, capable of providing unique and important insights into various factors that shape scientific reporting.

Concluding Remarks

This dissertation is a valuable asset to the literature due to its multifaceted approach to studying and identifying bias in published reports via more objective technology-driven tools such as GRADEpro and Topic Modeling. Further, this dissertation's thorough theoretical and conceptual treatment of bias adds a layer of complexity regarding *how* bias is perceived in today's academic climate.

Moving forward, researchers should work to continue studies of bias—especially as it relates to potential errors within their own work. More importantly, scholars should not be afraid to test tools such as GRADEpro and Topic Modeling, themselves. Once such tools are incorporated more into Public Health, Health Education, and Health Promotion discourse and research, we, as social scientists, can contribute our small part to reduce problematic practices hampering the quality of the scientific enterprise.

Though eliminating all bias from research is an impossible task, starting a dialogue and raising awareness of intentional and unintentional biases are ideal starting points. By bringing bias, and studies of bias, to the forefront of the scientific literature, and not relegating them to the commentary sections of academic journals, we stand to frame bias as a constantly evolving and complex issue that warrants continual vigilance. Further, as studies of bias increase, researchers may discover (or come to develop) *newer* tools and techniques to assist the scientific community with even better bias detection and mitigation beyond the tools presented in this dissertation. Overall, however, we should strive to uphold the integrity of science and the scientific enterprise. Science can promote much knowledge for the greater good. Undetected bias in science, however, stands to achieve exactly the opposite.

REFERENCES

Aarts, O. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716. https://doi.org/10.1126/science.aac4716

Abdelmalek M., Suzuki A., Guy C., Unalp A., Colvin R., Johnson R. (2010). Increased fructose consumption is associated with fibrosis severity in patients with nonalcoholic fatty liver disease. *Hepatology*, *51*(6), 1961–1971. https://doi.org/10.1002/hep.23535

Akhtar-Danesh, N., Baumann, A., & Cordingley, L. (2008). Q-methodology in nursing research: a promising method for the study of subjectivity. *Western Journal of Nursing Research*, *30*(6), 759–773. https://doi.org/10.1177/0193945907312979

Alberts, B. (2011). Science Breakthroughs. *Science*, *334*(6063), 1604–1604. https://doi.org/10.1126/science.1217831

Allison, P. Book Reviews. (2002). *British Journal of Mathematical and Statistical Psychology*, *55*(1), 193–196. https://doi.org/10.1348/000711002159653

Altman, D. (2002). Poor-Quality Medical Research: What Can Journals Do? *JAMA*, *287*(21), 2765–2767. https://doi.org/10.1001/jama.287.21.2765

Perrin, A., & Duggan, M. (2015, June 26). Americans' Internet Access: 2000-2015. Retrieved from http://www.pewinternet.org/2015/06/26/americans-internet-access-2000-2015

Anaya, L. H. (2011). *Comparing Latent Dirichlet Allocation and Latent Semantic Analysis as Classifiers* (Doctoral Dissertation). Retrieved from

https://digital.library.unt.edu/ark:/67531/metadc103284/m2/1/high_res_d/dissertation.pdf

Armstrong, D., Gosling, A., Weinman, J., & Marteau, T. (1997). The place of inter-rater reliability in qualitative research: An empirical study. *Sociology*, *31*(3), 597–606. https://doi.org/10.1177/0038038597031003015

Arnold, L., Lofthouse, N., & Hurt, E. (2012). Artificial food colors and attention-deficit/hyperactivity symptoms: Conclusions to dye for. *Neurotherapeutics*, *9*(3), 599–609. https://doi.org/10.1007/s13311-012-0133-x

Arrow, M., & Aronson, M. (2016, January 8). Seven culture-defining differences between UK and US ads. Retrieved April 29, 2018, from http://www.theguardian.com/media-network/2016/jan/08/culture-defining-differences-uk-us-ads

Artificial Dyes - American Chemical Society. (n.d.). Retrieved April 9, 2018, from https://www.acs.org/content/acs/en/education/resources/highschool/chemmatters/past-issues/2015-2016/october-2015/articial-dyes.html

Artino, A. R., & Naismith, L. M. (2015). 'But how do you really feel?' Measuring emotions in medical education research. *Medical Education*, *49*(2), 140–142. https://doi.org/10.1111/medu.12642

Åsberg, R., Hummerdal, D., & Dekker, S. (2011). There are no qualitative methods – nor quantitative for that matter: the misleading rhetoric of the qualitative–quantitative argument. *Theoretical Issues in Ergonomics Science*, *12*(5), 408–415. https://doi.org/10.1080/1464536X.2011.559292

Aschw, C., & King. (2015, August 19). Science Isn't Broken. Retrieved March 11, 2018, from https://fivethirtyeight.com/features/science-isnt-broken/

Atkins, D., Best, D., Briss, P., Eccles, M., Falck-Ytter, Y., Flottorp, S., … Al, E. (2004). Grading quality of evidence and strength of recommendations. *British Medical Journal*. 398(1490) https://doi.org/10.1136/bmj.328.7454.1490

Swinbum B., Caterson, I., Seidell, J., & James W. (2004). Diet, nutrition and the prevention of excess weight gain and obesity. *Public Health Nutrition*, *7*(1a), 123–146. https://doi.org/10.1079/PHN2003585

Babcock, Q., & Byrne, B. T. (n.d.). Student perceptions of methylphenidate abuse at a public liberal arts college. *J Am Coll Health*.

Baier, K. (1958). The moral point of view: A rational basis of ethics. *Philosophical Review*, *69*(4), 548–553.

Baker, M. (2016). Stat-checking software stirs up psychology. *Nature News*, *540*(7631), 151. https://doi.org/10.1038/540151a

Ball, P. (2017) It's not just you: science papers are getting harder to read. *Nature News*. https://doi.org/10.1038/nature.2017.21751

Banks, R., & Ford, R. (2008). (How) Does unconscious bias matter: Law, politics, and racial Inequality. *Emory Law Journal, 58*, 1053.

Barden, J., Derry, S., McQuay, H. J., & Moore, R. A. (2006). Bias from industry trial funding? A framework, a suggested approach, and a negative result. *Pain*, *121*(3), 207–218. https://doi.org/10.1016/j.pain.2005.12.011

Barry, A. E., Szucs, L. E., Reyes, J. V., Ji, Q., Wilson, K. L., & Thompson, B. (2016). failure to report effect sizes: The handling of quantitative results in published health education and behavior research. *Health Education & Behavior*, *43*(5), 518–527. https://doi.org/10.1177/1090198116669521

Benson, G. C. S. (1989). Codes of ethics. *Journal of Business Ethics*, *8*(5), 305–319. https://doi.org/10.1007/BF00381721

Bes-Rastrollo, M., Schulze, M., Ruiz-Canela, M., & Martinez-Gonzalez, M. (2013). Financial conflicts of interest and reporting bias regarding the association between sugar-sweetened beverages and weight gain: A systematic review of systematic reviews. *PLOS Medicine*, *10*(12), e1001578. https://doi.org/10.1371/journal.pmed.1001578

Blei, D., Ng, A., & Jordan, M. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, *3*(Jan), 993–1022.

Bohannon, J. (2013). Who's afraid of peer review? *Science*, *342*(6154), 60–65. https://doi.org/10.1126/science.342.6154.60

Bohannon, J. (2015). Many psychology papers fail replication test. *Science*, *349*(6251), 910–911. https://doi.org/10.1126/science.349.6251.910

Bollen, K. A., & Paxton, P. (1998). Detection and determinants of bias in subjective measures. *American Sociological Review*, *63*(3), 465–478. https://doi.org/10.2307/2657559

Boring, E. G. (1919). Mathematical vs. scientific significance. *Psychological Bulletin*, *16*(10), 335–338.

Bothwell, L., Greene, J., Podolsky, S., & Jones, D. (2016). Assessing the gold standard — Lessons from the history of RCTs. *New England Journal of Medicine*, *374*(22), 2175–2181. https://doi.org/10.1056/NEJMms1604593

Bowman, J. D. (2014). Predatory publishing, questionable peer review, and fraudulent conferences. *American Journal of Pharmaceutical Education*, *78*(10). https://doi.org/10.5688/ajpe7810176

Braunack-Mayer, A. J. (2001). What makes a problem an ethical problem? An empirical perspective on the nature of ethical problems in general practice. *Journal of Medical Ethics*, *27*(2), 98–103. https://doi.org/10.1136/jme.27.2.98

Breggin, P. (2007). *Talking back to ritalin: What doctors aren't telling you about stimulants and ADHD*. Da Capo Press, Incorporated.

Bricocoli, M., & Cucca, R. (2016). Social mix and housing policy: Local effects of a misleading rhetoric. The case of Milan. *Urban Studies*, *53*(1), 77–91. https://doi.org/10.1177/0042098014560499

Brinton, E. A., Eisenberg, S., & Breslow, J. L. (1990). A low-fat diet decreases high density lipoprotein (HDL) cholesterol levels by decreasing HDL apolipoprotein transport rates. *The Journal of Clinical Investigation*, *85*(1), 144–151. https://doi.org/10.1172/JCI114405

Brownell Kelly D., & Warner Kenneth E. (2009). The perils of ignoring history: Big tobacco played dirty and millions died. How similar is big food? *The Milbank Quarterly*, *87*(1), 259–294. https://doi.org/10.1111/j.1468-0009.2009.00555.x

Burgess, D., van Ryn, M., Dovidio, J., & Saha, S. (2007). Reducing racial bias among health care providers: Lessons from social-cognitive psychology. *Journal of General Internal Medicine*, *22*(6), 882–887. https://doi.org/10.1007/s11606-007-0160-1

Calder, B. J. (1977). Focus groups and the nature of qualitative marketing research. *Journal of Marketing Research*, *14*(3), 353–364. https://doi.org/10.2307/3150774

Calin-Jageman, R. J. (2017). After p-values: The new statistics for undergraduate neuroscience education. *Journal of Undergraduate Neuroscience Education*, *16*(1), E1–E4.

Casadevall, A., & Fang, F. C. (2012). Reforming science: Methodological and sultural reforms. *Infection and Immunity*, *80*(3), 891–896. https://doi.org/10.1128/IAI.06183-11

Chan, A., & Altman, D. (2005). Identifying outcome reporting bias in randomised trials on PubMed: review of publications and survey of authors. *BMJ (Clinical Research Ed.)*, *330*(7494), 753. https://doi.org/10.1136/bmj.38356.424606.8F

Chasela, C., Hudgens, M., Jamieson, D., Kayira, D., Hosseinipour, M., Kourtis, A., … van der Horst, C. (2010). Maternal or infant antiretroviral drugs to reduce HIV-1 transmission. *New England Journal of Medicine*, *362*(24), 2271–2281. https://doi.org/10.1056/NEJMoa0911486

Chin, J. (1990). Current and future dimensions of the HIV/AIDS pandemic in women and children. *The Lancet*, *336*(8709), 221–224. https://doi.org/10.1016/0140-6736(90)91743-T

Chiu, D., & Duesberg, P. (1995). The toxicity of azidothymidine (AZT) on human and animal cells in culture at concentrations used for antiviral therapy. *Genetica*, *95*(1–3), 103–109. https://doi.org/10.1007/BF01435004

Chopra, S. (2003). Industry funding of clinical Trials: benefit or bias? *JAMA*, *290*(1), 113–114. https://doi.org/10.1001/jama.290.1.113

Clifford, T. J., Barrowman, N. J., & Moher, D. (2002). Funding source, trial outcome and reporting quality: are they related? Results of a pilot study. *BMC Health Services Research*, *2*, 18. https://doi.org/10.1186/1472-6963-2-18

Cohen, J. (2011). HIV treatment as prevention. *Science*, *334*(6063), 1628–1628. https://doi.org/10.1126/science.334.6063.1628

Cokol, M., Ozbay, F., & Rodriguez-Esteban, R. (2008). Retraction rates are on the rise. *EMBO Reports*, *9*(1), 2. https://doi.org/10.1038/sj.embor.7401143

Collier, D., & Mahoney, J. (1996). Insights and pitfalls: Selection bias in qualitative research. *World Politics*, *49*(1), 56–91. https://doi.org/10.1353/wp.1996.0023

Collier, R. J., & Bauman, D. E. (2014). Update on human health concerns of recombinant bovine somatotropin use in dairy cows. *Journal of Animal Science*, *92*(4), 1800–1807. https://doi.org/10.2527/jas.2013-7383

Cook, B. G. (2014). A Call for examining replication and bias in special education research. *Remedial and Special Education*, *35*(4), 233–246. https://doi.org/10.1177/0741932514528995

Cooper, T. (2000). *Handbook of Administrative Ethics*. CRC Press.

Coote, B. (1987). The hunger crop: poverty and the sugar industry. *The Hunger Crop: Poverty and the Sugar Industry*. Retrieved from https://www.cabdirect.org/cabdirect/abstract/19886705695

Corlett, J. A. (2005). Ethical issues in journal peer-review. *Journal of Academic Ethics*, *2*(4), 355–366. https://doi.org/10.1007/s10805-005-9001-1

Corrigan, O. (2003). Empty ethics: the problem with informed consent. *Sociology of Health & Illness*, *25*(7), 768–792. https://doi.org/10.1046/j.1467-9566.2003.00369.x

Cox, D. J., Merkel, R. L., Kovatchev, B., & Seward, R. (2000). Effect of stimulant medication on driving performance of young adults with attention-deficit hyperactivity disorder: A preliminary double-blind placebo controlled trial. *The Journal of Nervous and Mental Disease*, *188*(4), 230.

Csada, R. D., James, P. C., & Espie, R. H. M. (1996). The "file drawer problem" of non-significant results: Does it apply to biological research? *Oikos*, *76*(3), 591–593. https://doi.org/10.2307/3546355

Curran, J. W., Jaffe, H. W., Hardy, A. M., Morgan, W. M., Selik, R. M., & Dondero, T. J. (1988). Epidemiology of HIV infection and AIDS in the United States. *Science*, *239*(4840), 610–616. https://doi.org/10.1126/science.3340847

Delaney, M. (2006). History of HAART – the true story of how effective multi-drug therapy was developed for treatment of HIV disease. *Retrovirology*, *3*(Suppl 1), S6. https://doi.org/10.1186/1742-4690-3-S1-S6

Delgado-Rodríguez, M., & Llorca, J. (2004). Bias. *Journal of Epidemiology & Community Health*, *58*(8), 635–641. https://doi.org/10.1136/jech.2003.008466

Dellarocas C., Zhang X., & Awad Neveen F. (2007). Exploring the value of online product reviews in forecasting sales: The case of motion pictures. *Journal of Interactive Marketing*, *21*(4), 23–45. https://doi.org/10.1002/dir.20087

Dennis, B. (2015, November 3). Nearly 60 percent of Americans — the highest ever — are taking prescription drugs. *Washington Post*. Retrieved from https://www.washingtonpost.com/news/to-your-health/wp/2015/11/03/more-americans-than-ever-are-taking-prescription-drugs/

Diederen, P. (1999). *Innovation and Research Policies: An International Comparative Analysis*. Edward Elgar Publishing.

Diener, E., & Biswas-Diener, R. (n.d.). The replication crisis in psychology. Retrieved March 11, 2018, from http://nobaproject.com/modules/the-replication-crisis-in-psychology

Dohoo, I. R., DesCôteaux, L., Leslie, K., Fredeen, A., Shewfelt, W., Preston, A., & Dowling, P. (2003). A meta-analysis review of the effects of recombinant bovine somatotropin. *Canadian Journal of Veterinary Research*, *67*(4), 252–264.

Donders, A. R. T., van der Heijden, G. J. M. G., Stijnen, T., & Moons, K. G. M. (2006). Review: a gentle introduction to imputation of missing values. *Journal of Clinical Epidemiology*, *59*(10), 1087–1091. https://doi.org/10.1016/j.jclinepi.2006.01.014

Drapeau, M. (2002). Subjectivity in research: Why not? But…. *The Qualitative Report*, *7*(3), 1–15.

Drewnowski, A., & Specter, S. E. (2004). Poverty and obesity: the role of energy density and energy costs. *The American Journal of Clinical Nutrition*, *79*(1), 6–16. https://doi.org/10.1093/ajcn/79.1.6

Earp, B. D., & Trafimow, D. (2015). Replication, falsification, and the crisis of confidence in social psychology. *Frontiers in Psychology*, *6*. https://doi.org/10.3389/fpsyg.2015.00621

Easterbrook, P. J., Gopalan, R., Berlin, J. A., & Matthews, D. R. (1991). Publication bias in clinical research. *The Lancet*, *337*(8746), 867–872. https://doi.org/10.1016/0140-6736(91)90201-Y

Egger, M., Zellweger-Zähner, T., Schneider, M., Junker, C., Lengeler, C., & Antes, G. (1997). Language bias in randomised controlled trials published in English and German. *The Lancet*, *350*(9074), 326–329. https://doi.org/10.1016/S0140-6736(97)02419-7

Epstein, S. (1996). *Impure Science: AIDS, Activism, and the Politics of Knowledge*. University of California Press.

Evan, W. M. (1962). Role Strain and the norm of reciprocity in research organizations. *American Journal of Sociology*, *68*(3), 346–354. https://doi.org/10.1086/223354

Fairclough, N. (2013). *Critical Discourse Analysis: The Critical Study of Language*. Routledge.

Fan, X., & Thompson, B. (2001). Confidence intervals for effect sizes: Confidence intervals about score reliability coefficients, please: An EPM guidelines editorial. *Educational*

and Psychological Measurement, *61*(4), 517–531. https://doi.org/10.1177/0013164401614001

Fanelli, D. (2009). How many scientists fabricate and falsify research? A systematic review and meta-analysis of survey data. *PLOS ONE*, *4*(5), e5738. https://doi.org/10.1371/journal.pone.0005738

Fang, F. C., Steen, R. G., & Casadevall, A. (2012). Misconduct accounts for the majority of retracted scientific publications. *Proceedings of the National Academy of Sciences*, *109*(42), 17028–17033. https://doi.org/10.1073/pnas.1212247109

Field, A., Miles, J., & Field, Z. (2012). *Discovering statistics using R*. SAGE Publications.

Fielden, T. (2017). Most Scientist Can't Replicate Studies By Their Peers. *British Broadcasting Corporation*. Retrieved from https://www.bbc.com/news/science-environment-39054778

Fischl, M. A., Richman, D. D., Grieco, M. H., Gottlieb, M. S., Volberding, P. A., Laskin, O. L., … Group, T. A. C. W. (1987). The efficacy of azidothymidine (AZT) in the treatment of patients with AIDS and AIDS-related complex. *New England Journal of Medicine*, *317*(4), 185–191. https://doi.org/10.1056/NEJM198707233170401

Fischler, C. (1987). Attitudes towards sugar and sweetness in historical and social perspective. In *Sweetness* (pp. 83–98). Springer, London. https://doi.org/10.1007/978-1-4471-1429-1_6

Flash C., Landovitz R., Mera Giler R., Ng L., Magnuson D., Bush Wooley S., & Rawlings K. (2014). Two years of Truvada for pre-exposure prophylaxis utilization in the US.

Journal of the International AIDS Society, *17*(4S3), 19730. https://doi.org/10.7448/IAS.17.4.19730

Fogoros, R. N., & MD. (n.d.). What Are BHA and BHT? Are They Safe? Retrieved April 15, 2018, from https://www.verywellfit.com/bha-and-bht-keep-foods-fresh-but-are-they-safe-2506579

Food coloring is bad for us, but the FDA won't admit that. (n.d.). Retrieved April 9, 2018, from http://www.slate.com/articles/health_and_science/science/2016/07/food_coloring_is_bad_for_us_but_the_fda_won_t_admit_that.html

Galea, J. T., Kinsler, J. J., Salazar, X., Lee, S.-J., Giron, M., Sayles, J. N., … Cunningham, W. E. (2011). Acceptability of Pre-Exposure Prophylaxis (PrEP) as an HIV prevention strategy: Barriers and facilitators to PrEP uptake among at-risk Peruvian populations. *International Journal of STD & AIDS*, *22*(5), 256–262. https://doi.org/10.1258/ijsa.2009.009255

Garrow, D. J. (2015). *Liberty and sexuality: The right to privacy and the making of roe v. wade*. Open Road Media.

Gawande, A. (2016, June 10). The mistrust of science. *The New Yorker*. Retrieved from https://www.newyorker.com/news/news-desk/the-mistrust-of-science

Office of the Surgeon General. (2004). *The health consequences of smoking: A report from the Surgeon General*. Centers for Disease Control and Prevention (US). Retrieved from https://www.ncbi.nlm.nih.gov/books/NBK44699/

Gentilviso, C. (2010). The 50 Worst Inventions. *Time*. Retrieved from http://content.time.com/time/specials/packages/article/0,28804,1991915_1991909_1991785,00.html

Gibson, A. A., & Sainsbury, A. (2017). Strategies to improve adherence to dietary weight loss interventions in research and real-world settings. *Behavioral Sciences*, *7*(3). https://doi.org/10.3390/bs7030044

Gino, F., & Bazerman, M. H. (2009). When misconduct goes unnoticed: The acceptability of gradual erosion in others' unethical behavior. *Journal of Experimental Social Psychology*, *45*(4), 708–719. https://doi.org/10.1016/j.jesp.2009.03.013

Glasziou, P., Vandenbroucke, J., & Chalmers, I. (2004). Assessing the quality of research. *BMJ : British Medical Journal*, *328*(7430), 39–41.

Glover, T., & Mitchell, K. (2015). *An Introduction to Biostatistics: Third Edition*. Waveland Press.

Gordon, K. R. (2006). Objectivity in science. *Science*, *311*(5765), 1240–1241. https://doi.org/10.1126/science.311.5765.1240c

Gorsuch, R. L. (1988). Exploratory factor analysis. In *handbook of multivariate experimental psychology* (pp. 231–258). Springer, Boston, MA. https://doi.org/10.1007/978-1-4613-0893-5_6

Gottschalk, S. A., & Mafael, A. (2017). Cutting through the online review jungle — investigating selective eWOM processing. *Journal of Interactive Marketing*, *37*, 89–104. https://doi.org/10.1016/j.intmar.2016.06.001

Grady, D. (2012, May 14). Taking truvada to prevent H.I.V. also comes with risks. *The New York Times*. Retrieved from https://www.nytimes.com/2012/05/15/health/policy/taking-truvada-to-prevent-hiv-also-comes-with-risks.html

Griffiths, T. (2002). *Gibbs sampling in the generative model of Latent Dirichlet Allocation*.

Grunert, K. G., & Wills, J. M. (2007). A review of European research on consumer response to nutrition information on food labels. *Journal of Public Health*, *15*(5), 385–399. https://doi.org/10.1007/s10389-007-0101-9

Gupta, A. (2013). Fraud and misconduct in clinical research: A concern. *Perspectives in Clinical Research*, *4*(2), 144–147. https://doi.org/10.4103/2229-3485.111800

Gupta, J., & van der Zaag, P. (2009). The politics of water science: On unresolved water problems and biased research agendas. *Global Environmental Politics*, *9*, 14–23. https://doi.org/10.1162/glep.2009.9.2.14

Guyatt, G. H., Oxman, A. D., Kunz, R., Vist, G. E., Falck-Ytter, Y., & Schünemann, H. J. (2008). What is "quality of evidence" and why is it important to clinicians? *BMJ : British Medical Journal*, *336*(7651), 995–998. https://doi.org/10.1136/bmj.39490.551019.BE

Hadorn, G. H., Biber-Klemm, S., Grossenbacher-Mansuy, W., Hoffmann-Riem, H., Joye, D., Pohl, C., … Zemp, E. (2008). The emergence of transdisciplinarity as a form of research. In *Handbook of Transdisciplinary Research* (pp. 19–39). Springer, Dordrecht. https://doi.org/10.1007/978-1-4020-6699-3_2

Hammersley, M., & Gomm, R. (1997). Bias in social research. *Sociological Research Online*, *2*. https://doi.org/10.5153/sro.55

Hanson, K. O. (2014). Six unavoidable ethical dilemmas every professional faces. *Business and Society Review*, *119*(4), 537–552. https://doi.org/10.1111/basr.12045

Hayat, M. J., Powell, A., Johnson, T., & Cadwell, B. L. (2017). Statistical methods used in the public health literature and implications for training of public health professionals. *PLoS ONE*, *12*(6). https://doi.org/10.1371/journal.pone.0179032

Head, M. L., Holman, L., Lanfear, R., Kahn, A. T., & Jennions, M. D. (2015). The extent and consequences of p-hacking in science. *PLOS Biology*, *13*(3), e1002106. https://doi.org/10.1371/journal.pbio.1002106

Henderson, M. (2010). Problems with peer review. *BMJ*, *340*, c1409. https://doi.org/10.1136/bmj.c1409

Hesselmann, F., Graf, V., Schmidt, M., & Reinhart, M. (2017). The visibility of scientific misconduct: A review of the literature on retracted journal articles. *Current Sociology*, *65*(6), 814–845. https://doi.org/10.1177/0011392116663807

Hoffman, M., Bach, F. R., & Blei, D. M. (2010). Online Learning for Latent Dirichlet Allocation. In J. D. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. S. Zemel, & A. Culotta (Eds.), *Advances in Neural Information Processing Systems 23* (pp. 856–864). Curran Associates, Inc. Retrieved from http://papers.nips.cc/paper/3902-online-learning-for-latent-dirichlet-allocation.pdf

Holmes, D. (2012). FDA paves the way for pre-exposure HIV prophylaxis. *The Lancet*, *380*(9839), 325. https://doi.org/10.1016/S0140-6736(12)61235-5

Hu, Y., Boyd-Graber, J., Satinoff, B., & Smith, A. (2014). Interactive topic modeling. *Machine Learning*, *95*(3), 423–469. https://doi.org/10.1007/s10994-013-5413-0

Hubbard, R. (2015). *Corrupt Research: The Case for Reconceptualizing Empirical Management and Social Science*. SAGE Publications.

Ciadianni, R.B. (2001.). *Influence: Science and Practice*. Fourth Edition. Retrieved from https://www.influenceatwork.com/wp-content/uploads/2012/02/Influence_SP.pdf

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

Jellinek, M. S. (2003). Psychiatric drugs and children. *Science*, *300*(5621), 901–901. https://doi.org/10.1126/science.300.5621.901b

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science*, *23*(5), 524–532. https://doi.org/10.1177/0956797611430953

Jones, M. C. (2014). *Endangered languages and new technologies*. Cambridge University Press.

Juul-Moller, S., Edvardsson, N., Sorensen, S., Jahnmatz, B., Rosén, A., & Omblus, R. (1992). Double-blind trial of aspirin in primary prevention of myocardial infarction in patients with stable chronic angina pectoris. *The Lancet*, *340*(8833), 1421–1425. https://doi.org/10.1016/0140-6736(92)92619-Q

Kashyap, U. N., Gupta, V., & Raghun, H. V. (2013). Comparison of drug approval process in United States and Europe. *Pharmaceutical Science & Resolutions.* 56(6) 131-136

Kavanagh, B. P. (2009). The GRADE system for rating clinical guidelines. *PLOS Medicine*, *6*(9), e1000094. https://doi.org/10.1371/journal.pmed.1000094

Kearns, C. E., Schmidt, L. A., & Glantz, S. A. (2016). Sugar industry and coronary heart disease research: A historical analysis of internal industry documents. *JAMA Internal Medicine*, *176*(11), 1680–1685. https://doi.org/10.1001/jamainternmed.2016.5394

King, G. (2011). The Social Science Data Revolution. Power Point. Retrieved from https://gking.harvard.edu/files/gking/files/evbase-horizonsp.pdf

Lakeh, A. B., & Ghaffarzadegan, N. (2017). Global trends and regional variations in studies of HIV/AIDS. *Scientific Reports*, *7*(1), 4170. https://doi.org/10.1038/s41598-017-04527-6

Lakoff, R. (1992). *Talking power: The politics of language*. Basic Books, LLC.

Lambdin, C. (2012). Significance tests as sorcery: Science is empirical—significance tests are not. *Theory & Psychology*, *22*(1), 67–90. https://doi.org/10.1177/0959354311429854

Larsen, P. O., & von Ins, M. (2010). The rate of growth in scientific publication and the decline in coverage provided by Science Citation Index. *Scientometrics*, *84*(3), 575–603. https://doi.org/10.1007/s11192-010-0202-z

Larson, R. C., Ghaffarzadegan, N., & Xue, Y. (2014). Too pany PhD graduates or too few academic job openings: The basic reproductive number R0 in academia. *Systems Research and Behavioral Science*, *31*(6), 745–750. https://doi.org/10.1002/sres.2210

Lee, D. K. (2016). Alternatives to P value: confidence interval and effect size. *Korean Journal of Anesthesiology*, *69*(6), 555–562. https://doi.org/10.4097/kjae.2016.69.6.555

Lee, M.-H., Wu, Y.-T., & Tsai, C.-C. (2009). Research trends in science education from 2003 to 2007: A content analysis of publications in selected journals. *International Journal of Science Education*, *31*(15), 1999–2020. https://doi.org/10.1080/09500690802314876

Leibowitz, A. A., Parker, K. B., & Rotheram-Borus, M. J. (2011). A US policy perspective on oral preexposure prophylaxis for HIV. *American Journal of Public Health*, *101*(6), 982–985. https://doi.org/10.2105/AJPH.2010.300066

Levine, T. R., Asada, K. J., & Carpenter, C. (2009). Sample sizes and effect sizes are negatively correlated in meta-analyses: Evidence and implications of a publication bias against non-significant findings. *Communication Monographs*, *76*(3), 286–302. https://doi.org/10.1080/03637750903074685

Lexchin, J. (2012a). Sponsorship bias in clinical research. *The International Journal of Risk & Safety in Medicine*, *24*(4), 233–242. https://doi.org/10.3233/JRS-2012-0574

Lexchin, J. (2012b). Those who have the gold make the evidence: how the pharmaceutical industry biases the outcomes of clinical trials of medications. *Science and Engineering Ethics*, *18*(2), 247–261. https://doi.org/10.1007/s11948-011-9265-3

Lexchin, J., Bero, L. A., Djulbegovic, B., & Clark, O. (2003). Pharmaceutical industry sponsorship and research outcome and quality: systematic review. *BMJ : British Medical Journal*, *326*(7400), 1167.

Linell, P. (2004). *The Written Language Bias in Linguistics: Its Nature, Origins and Transformations*. Routledge Press.

Liu, L., Tang, L., Dong, W., Yao, S., & Zhou, W. (2016). An overview of topic modeling and its current applications in bioinformatics. *SpringerPlus*, *5*(1). https://doi.org/10.1186/s40064-016-3252-8

Loken, E., & Gelman, A. (2017). Measurement error and the replication crisis. *Science*, *355*(6325), 584–585. https://doi.org/10.1126/science.aal3618

Long, L. (2013). *Understanding GRADE: An Introduction*. Power Point. Retrieved February 15, 2018, from https://medicine.exeter.ac.uk/media/universityofexeter/medicalschool/events/docs/Oct_2013_GRADE.pdf

Lyles, A. (2002). Direct marketing of pharmaceuticals to consumers. *Annual review of public health*, *23*(1), 73–91. https://doi.org/10.1146/annurev.publhealth.23.100901.140537

Maffini, M. V., Rubin, B. S., Sonnenschein, C., & Soto, A. M. (2006). Endocrine disruptors and reproductive health: The case of bisphenol-A. *Molecular and Cellular Endocrinology*, *254–255*, 179–186. https://doi.org/10.1016/j.mce.2006.04.033

Maheswaran, D., & Chen, C. Y. (2006). Nation equity: Incidental emotions in country-of-origin effects. *Journal of Consumer Research*, *33*(3), 370–376. https://doi.org/10.1086/508521

Malik, V. S., Schulze, M. B., & Hu, F. B. (2006). Intake of sugar-sweetened beverages and weight gain: a systematic review–. *The American Journal of Clinical Nutrition*, *84*(2), 274–288. https://doi.org/10.1093/ajcn/84.1.274

Marazzi, M. C., Germano, P., Liotta, G., Buonomo, E., Guidotti, G., & Palombi, L. (2006). Pediatric highly active antiretroviral therapy in Mozambique: an integrated model of care. *Minerva Pediatrica*, *58*(5), 483–490.

Marcoulides, G. A. (1998). *Modern Methods for Business Research*. Psychology Press.

Marcus, A., & Oransky, I. (2014). What studies of retractions tell us. *Journal of Microbiology & Biology Education*, *15*(2), 151–154. https://doi.org/10.1128/jmbe.v15i2.855

Marcus R., Ferrand R.A., Kranzer K., & Bekker L.G. (2017). The case for viral load testing in adolescents in resource-limited settings. *Journal of the International AIDS Society*, *20*(S7), e25002. https://doi.org/10.1002/jia2.25002

Marsh, H. W., Jayasinghe, U. W., & Bond, N. W. (2008). Improving the peer-review process for grant applications: reliability, validity, bias, and generalizability. *American Psychologist*, *63*(3), 160–168.

Martin Leslie R., & Friedman Howard S. (2001). Comparing personality scales across time: An illustrative study of validity and consistency in life-span archival data. *Journal of Personality*, *68*(1), 85–110. https://doi.org/10.1111/1467-6494.00092

Mason, S. (2016). Special issue on: HIV across the lifespan. *Journal of HIV/AIDS & Social Services*, *15*(4), 349–352. https://doi.org/10.1080/15381501.2016.1239976

Matic, S., Lazarus, J. V., & Donoghoe, M. C. (2006). HIV/AIDS in Europe: moving from death sentence to chronic disease management. *HIV/AIDS in Europe: Moving from Death Sentence to Chronic Disease Management*. Retrieved from https://www.cabdirect.org/cabdirect/abstract/20063030422

McArdle, M. (2011, August 31). How Bias Works. *The Atlantic*. Retrieved from https://www.theatlantic.com/business/archive/2011/08/how-bias-works/244393/

McCann, D., Barrett, A., Cooper, A., Crumpler, D., Dalen, L., Grimshaw, K., … Stevenson, J. (2007). Food additives and hyperactive behaviour in 3-year-old and 8/9-year-old children in the community: a randomised, double-blinded, placebo-controlled trial. *Lancet (London, England), 370*(9598), 1560–1567. https://doi.org/10.1016/S0140-6736(07)61306-3

McGarity, T. O., & Wagner, W. E. (2008). *Bending science: How special interests corrupt public health research*. Harvard Univ Press.

MD, M. C., Yifrah Kaminer MD, M., & MD, A. M. (2002). Megadose intranasal methylphenidate (Ritalin) abuse in adult attention deficit hyperactivity disorder. *Substance Abuse, 23*(3), 165–169. https://doi.org/10.1080/08897070209511486

Meader, N., King, K., Llewellyn, A., Norman, G., Brown, J., Rodgers, M., … Stewart, G. (2014). A checklist designed to aid consistency and reproducibility of GRADE assessments: development and pilot validation. *Systematic Reviews, 3*, 82. https://doi.org/10.1186/2046-4053-3-82

Meijer, W. M., Faber, A., Ban, E. van den, & Tobi, H. (2009). Current issues around the pharmacotherapy of ADHD in children and adults. *Pharmacy World & Science, 31*(5), 509–516. https://doi.org/10.1007/s11096-009-9302-3

Kravitz, M. (2017). 6 foods that are legal in the US but banned in other countries. Retrieved April 8, 2018, from http://www.businessinsider.com/foods-illegal-outside-us-2017-3

Michaels, D. (2008a). *Doubt is their product: How industry's assault on science threatens your health*. Oxford University Press, USA.

Michaels, D. (2008b, July 15). *It's not the answers that are biased, it's the questions.* Retrieved from http://www.washingtonpost.com/wp-dyn/content/article/2008/07/14/AR2008071402145.html

National Public Radio (NPR). (2009) *More students turning illegally to "smart" drugs.* Retrieved April 20, 2018, from https://www.npr.org/templates/story/story.php?storyId=100254163

Morton, W. A., & Stockton, G. G. (2000). Methylphenidate abuse and psychiatric side effects. *Primary care companion to the journal of clinical psychiatry*, *2*(5), 159–164.

Mullins, T. L. K., Lally, M., Zimet, G., & Kahn, J. A. (2015). Clinician attitudes toward CDC interim pre-exposure prophylaxis (PrEP) guidance and operationalizing PrEP for adolescents. *AIDS Patient Care and STDs*, *29*(4), 193–203. https://doi.org/10.1089/apc.2014.0273

Munafò, M. R., Nosek, B. A., Bishop, D. V. M., Button, K. S., Chambers, C. D., Percie du Sert, N., … Ioannidis, J. P. A. (2017). A manifesto for reproducible science. *Nature Human Behaviour*, *1*(1), 0021. https://doi.org/10.1038/s41562-016-0021

Munroe, R., Kaiser, J., & Malakoff, D. (2009). How much science is there? *Science*, *342*(6154), 58–59.

Noble, H., & Smith, J. (2015). Issues of validity and reliability in qualitative research. *Evidence-Based Nursing*, ebnurs-2015-102054. https://doi.org/10.1136/eb-2015-102054

Nuijten, M. B., Assen, M. A. L. M. van, Hartgerink, C. H. J., Epskamp, S., & Wicherts, J. (2017). The validity of the tool "statcheck" in discovering statistical reporting inconsistencies. *PsyArXiv*. https://doi.org/10.17605/OSF.IO/TCXAJ

Nuijten, M. B., Hartgerink, C. H. J., Assen, M. A. L. M. van, Epskamp, S., & Wicherts, J. M. (2016). The prevalence of statistical reporting errors in psychology (1985–2013). *Behavior Research Methods*, *48*(4), 1205–1226. https://doi.org/10.3758/s13428-015-0664-2

Nuzzo, R. (2014). Scientific method: statistical errors. *Nature*, *506*(7487), 150–152. https://doi.org/10.1038/506150a

Coffman & Odlyzko (2001) Growth of the internet. *AT&T Labs- Research*. Retrieved from http://www.dtc.umn.edu/~odlyzko/doc/oft.internet.growth.pdf

Ooms, M. E., Roos, J. C., Bezemer, P. D., Vijgh, V. D., J, W., Bouter, L. M., & Lips, P. (1995). Prevention of bone loss by vitamin D supplementation in elderly women: a randomized double-blind trial. *The Journal of Clinical Endocrinology & Metabolism*, *80*(4), 1052–1058. https://doi.org/10.1210/jcem.80.4.7714065

Ortega, A., & Navarrete, G. (2017). Bayesian hypothesis testing: An alternative to null hypothesis significance testing (NHST) in psychology and social sciences. *Bayesian Inference*. https://doi.org/10.5772/intechopen.70230

Ortlipp, M. (2008). Keeping and using reflective journals in the qualitative research Process. *The Qualitative Report*, *13*(4), 695–705.

Outwater, J. L., Nicholson, A., & Barnard, N. (1997). Dairy products and breast cancer: the IGF-I, estrogen, and bGH hypothesis. *Medical Hypotheses*, *48*(6), 453–461. https://doi.org/10.1016/S0306-9877(97)90110-9

Pannucci, C. J., & Wilkins, E. G. (2010). Identifying and avoiding bias in research. *Plastic and Reconstructive Surgery*, *126*(2), 619–625. https://doi.org/10.1097/PRS.0b013e3181de24bc

Park, S. (2014). Consumption of sugar-sweetened beverages among US adults in 6 states: Behavioral risk factor surveillance system. *Preventing Chronic Disease*, *11*. https://doi.org/10.5888/pcd11.130304

Pashler, H., & Harris, C. R. (2012). Is the Replicability Crisis Overblown? Three Arguments Examined. *Perspectives on Psychological Science*, *7*(6), 531–536. https://doi.org/10.1177/1745691612463401

Pashler, H., & Wagenmakers, E.-J. (2012). Editors' introduction to the special section on replicability in psychological science: A crisis of confidence? *Perspectives on Psychological Science*, *7*(6), 528–530. https://doi.org/10.1177/1745691612465253

Pechey, R., & Monsivais, P. (2016). Socioeconomic inequalities in the healthiness of food choices: Exploring the contributions of food expenditures. *Preventive Medicine*, *88*, 203–209. https://doi.org/10.1016/j.ypmed.2016.04.012

Péladeau, N., & Davoodi, E. (n.d.). Comparison of latent dirichlet modeling and factor analysis for topic extraction: A Lesson of History, 9. https://doi.org/10.24251/HICSS.2018.078.

Pickett, A. C., Valdez, D., & Barry, A. E. (2017). Psychometrics matter in health behavior: A long-term reliability generalization study. *American Journal of Health Behavior*, *41*(5), 544–552. https://doi.org/10.5993/AJHB.41.5.3

PMC, Europe. (2012). *Guidance on Pre-Exposure Oral Prophylaxis (PrEP) for Serodiscordant Couples, Men and Transgender Women Who Have Sex with Men at High Risk of HIV: Recommendations for Use in the Context of Demonstration Projects*. World Health Organization, Geneva. Retrieved from http://europepmc.org/abstract/med/23586123

Popay, J., & Williams, G. (1996). Public health research and lay knowledge. *Social Science & Medicine*, *42*(5), 759–768. https://doi.org/10.1016/0277-9536(95)00341-X

Pope, C., & Mays, N. (1995). Qualitative research: reaching the parts other methods cannot reach: an introduction to qualitative methods in health and health services research. *BMJ*, *311*(6996), 42–45. https://doi.org/10.1136/bmj.311.6996.42

Potter, N. N., & Hotchkiss, J. H. (2012). *Food Science: Fifth Edition*. Springer Science & Business Media.

PrEP | HIV Basics | HIV/AIDS | CDC. (2018, April 25). Retrieved May 21, 2018, from https://www.cdc.gov/hiv/basics/prep.html

PrEPWatch.Org (2012) PrEP efficacy trial results. Retrieved from https://www.prepwatch.org/wp-content/uploads/2017/05/PrEP_efficacy_results.pdf

Carroll, J.B. (1985). Exploratory factor analysis: A tutorial. *Current Topics in Human Intelligence, 1* , 25-58

Puhan, M. A., Schünemann, H. J., Murad, M. H., Li, T., Brignardello-Petersen, R., Singh, J. A., … GRADE Working Group. (2014). A GRADE Working Group approach for rating the quality of treatment effect estimates from network meta-analysis. *BMJ (Clinical Research Ed.)*, *349*, g5630.

Pyshcological Science Agenda News (2016) P-values under question. Retrieved March 11, 2018, from http://www.apa.org/science/about/psa/2016/03/p-values.aspx

Ransohoff, D. F. (2005). Bias as a threat to the validity of cancer molecular-marker research. *Nature Reviews. Cancer*, *5*(2), 142–149. https://doi.org/10.1038/nrc1550

Resnik, N. (n.d.). What is Ethics in Research & Why is it Important? Retrieved March 11, 2018, from https://www.niehs.nih.gov/research/resources/bioethics/whatis/index.cfm

Resources for Research Ethics Education. (n.d.). Retrieved March 11, 2018, from http://research-ethics.org/

Robison, L. S., Ananth, M., Hadjiargyrou, M., Komatsu, D. E., & Thanos, P. K. (2017). Chronic oral methylphenidate treatment reversibly increases striatal dopamine transporter and dopamine type 1 receptor binding in rats. *Journal of Neural Transmission*, *124*(5), 655–667. https://doi.org/10.1007/s00702-017-1680-4

Rodearmel, S. J., Wyatt, H. R., Stroebele, N., Smith, S. M., Ogden, L. G., & Hill, J. O. (2007). Small changes in dietary sugar and physical activity as an approach to preventing excessive weight gain: The America on the move family study. *Pediatrics*, *120*(4), e869–e879. https://doi.org/10.1542/peds.2006-2927

Rosen, S., & Fox, M. P. (2011). Retention in HIV care between testing and treatment in sub-saharan Africa: A Systematic Review. *PLOS Medicine*, *8*(7), e1001056. https://doi.org/10.1371/journal.pmed.1001056

Rosnow, R., & Rosenthal, R. (1989). Statisitcal procedures and the justification of knowledge in psychological science. *American Psychologist*, *44*(10), 1276.

Rychetnik, L., Frommer, M., Hawe, P., & Shiell, A. (2002). Criteria for evaluating evidence on public health interventions. *Journal of Epidemiology & Community Health*, *56*(2), 119–127. https://doi.org/10.1136/jech.56.2.119

Sackett, D. L. (1979). Bias in analytic research. *Journal of Chronic Diseases*, *32*(1), 51–63. https://doi.org/10.1016/0021-9681(79)90012-2

Safer, D. J., Zito, J. M., & Fine, E. M. (1996). Increased methylphenidate usage for attention deficit disorder in the 1990s. *Pediatrics*, *98*(6), 1084–1088.

Sanghavi, P., Jena, A. B., Newhouse, J. P., & Zaslavsky, A. M. (2015). Outcomes of basic versus advanced life support for out-of-hospital medical emergencies. *Annals of Internal Medicine*, *163*(9), 681. https://doi.org/10.7326/M15-0557

Sarniak. (2015). 9 types of research bias and how to avoid them | Articles | Quirks.com. Retrieved March 11, 2018, from https://www.quirks.com/articles/9-types-of-research-bias-and-how-to-avoid-them

Scharff, D. P., Mathews, K. J., Jackson, P., Hoffsuemmer, J., Martin, E., & Edwards, D. (2010). More than Tuskegee: Understanding mistrust about research participation. *Journal of Health Care for the Poor and Underserved*, *21*(3), 879–897. https://doi.org/10.1353/hpu.0.0323

School of Education at Johns Hopkins University-Center for Research and Reform. (n.d.). Retrieved March 11, 2018, from http://archive.education.jhu.edu/research/crre/index.html

Schulze, M. B., Manson, J. E., Ludwig, D. S., Colditz, G. A., Stampfer, M. J., Willett, W. C., & Hu, F. B. (2004). Sugar-sweetened beverages, weight gain, and incidence of Type 2 diabetes in young and middle-aged Women. *JAMA*, *292*(8), 927–934. https://doi.org/10.1001/jama.292.8.927

Shillington, A. M., Reed, M. B., Lange, J. E., Clapp, J. D., & Henry, S. (2006). College Undergraduate ritalin abusers in southwestern California: Protective and risk factors. *Journal of Drug Issues*, *36*(4), 999–1014. https://doi.org/10.1177/002204260603600411

Sica, A., Schioppa, T., Mantovani, A., & Allavena, P. (2006). Tumour-associated macrophages are a distinct M2 polarised population promoting tumour progression: potential targets of anti-cancer therapy. *European Journal of Cancer (Oxford, England: 1990)*, *42*(6), 717–727. https://doi.org/10.1016/j.ejca.2006.01.003

Šimundić, A.-M. (2013). Bias in research. *Biochemia Medica*, *23*(1), 12–15. https://doi.org/10.11613/BM.2013.003

Singh, J. A., & Mills, E. J. (2005). The abandoned trials of pre-exposure prophylaxis for HIV: What went wrong? *PLOS Medicine*, *2*(9), e234. https://doi.org/10.1371/journal.pmed.0020234

Smith, R. (1997). Peer review: reform or revolution? *BMJ : British Medical Journal*, *315*(7111), 759–760.

Song, F., Altman, D. G., Glenny, A.-M., & Deeks, J. J. (2003). Validity of indirect comparison for estimating efficacy of competing interventions: empirical evidence from published meta-analyses. *BMJ*, *326*(7387), 472. https://doi.org/10.1136/bmj.326.7387.472

Soumerai, S., & Koppel, R. (2017, June 7). Perspective. How bad science can lead to bad science journalism — and bad policy. *Washington Post*. Retrieved from https://www.washingtonpost.com/posteverything/wp/2017/06/07/how-bad-science-can-lead-to-bad-science-journalism-and-bad-policy/

Spruance, S. L., Reid, J. E., Grace, M., & Samore, M. (2004). Hazard ratio in clinical trials. *Antimicrobial Agents and Chemotherapy*, *48*(8), 2787–2792. https://doi.org/10.1128/AAC.48.8.2787-2792.2004

Srivastava, A. N., & Sahami, M. (2009). *Text Mining: Classification, Clustering, and Applications*. CRC Press.

Staats, C., & Paton, c. (n.d.). State of Science: Impllicit Bias Review.

Steen, R. G., Casadevall, A., & Fang, F. C. (2013). Why has the number of scientific retractions increased? *PLOS ONE*, *8*(7), e68397. https://doi.org/10.1371/journal.pone.0068397

Stigbrand, T. (2017). Retraction note to multiple articles in Tumor Biology. *Tumor Biology*, 1–6. https://doi.org/10.1007/s13277-017-5487-6

Stoker, D. (1995). Book Reviews : Daniel, H. D. Guardians of science: fairness and reliability of peer review. Translated by William E. Russey. 1993, Weinheim; New

York: VCH, 118pp, ISBN 3-527-29041-9. *Journal of Librarianship and Information Science*, *27*(2), 116–117. https://doi.org/10.1177/096100069502700210

Stroebe, W., & Strack, F. (2014). The alleged crisis and the illusion of exact replication. *Perspectives on Psychological Science*, *9*(1), 59–71. https://doi.org/10.1177/1745691613514450

Sugarman, J., & Mayer, K. H. (2013). Ethics and pre-exposure prophylaxis for HIV infection. *Journal of Acquired Immune Deficiency Syndromes (1999)*, *63 Suppl 2*, S135-139. https://doi.org/10.1097/QAI.0b013e3182987787

Sullivan, G. M., & Feinn, R. (2012). Using effect size—or why the p-value is not enough. *Journal of Graduate Medical Education*, *4*(3), 279–282. https://doi.org/10.4300/JGME-D-12-00156.1

Suominen, A., & Toivanen, H. (2016). Map of science with topic modeling: Comparison of unsupervised learning and human-assigned subject classification. *Journal of the Association for Information Science and Technology*, *67*(10), 2464–2476. https://doi.org/10.1002/asi.23596

Swaminathan, N., & Swaminathan, N. (n.d.). Use it or lose it: Why language changes over Time. Retrieved April 29, 2018, from https://www.scientificamerican.com/article/use-it-or-lose-it-why-lan/

Swanson, J. M., & Kinsbourne, M. (1980). Food dyes impair performance of hyperactive children on a laboratory learning test. *Science*, *207*(4438), 1485–1487. https://doi.org/10.1126/science.7361102

Thompson, B. (2002a). *Score reliability: Contemporary thinking on reliability issues*. SAGE Publications.

Thompson, B. (2002b). "Statistical," "practical," and "clinical": How many kinds of significance so counselors need to consider? *Journal of Counseling & Development*, *80*(1), 64–71. https://doi.org/10.1002/j.1556-6678.2002.tb00167.x

Thompson, B. (2006). *Foundations of Behavioral Statistics: An Insight-Based Approach*. Guilford Press.

Thompson, B. (2007). Effect sizes, confidence intervals, and confidence intervals for effect sizes. *Psychology in the Schools*, *44*(5), 423–432. https://doi.org/10.1002/pits.20234

Thompson, B. (2015). The case for using the general linear model as a unifying conceptual framework for teaching statistics and psychometric theory. *Journal of Methods and Measurement in the Social Sciences*, *6*(2), 30–41. https://doi.org/10.2458/v6i2.18801

Thompson, B., & Vacha-Haase, T. (2000). Psychometrics is datametrics: The test is not reliable. *Educational and Psychological Measurement*, *60*(2), 174–195. https://doi.org/10.1177/0013164400602002

Titov, I., & McDonald, R. (2008). Modeling Online Reviews with Multi-grain Topic Models. *ArXiv:0801.1063 [Cs]*. Retrieved from http://arxiv.org/abs/0801.1063

Trafimow, D., & Marks, M. (2015). Editorial. *Basic and Applied Social Psychology*, *37*(1), 1–2. https://doi.org/10.1080/01973533.2015.1012991

Trusty, J., Thompson, B., & Petrocelli, J. V. (2004). Practical guide for reporting effect size in quantitative research in the journal of counseling & development. *Journal of*

Counseling & Development, *82*(1), 107–110. https://doi.org/10.1002/j.1556-6678.2004.tb00291.x

Truth from the Dairy Aisle: Is Milk from Cows Receiving rbST Safe for my Family? (2012, March 19). Retrieved April 15, 2018, from https://scienceofmom.com/2012/03/19/truth-from-the-dairy-aisle-is-milk-from-cows-receiving-rbst-safe-for-my-family/

Tunis, S. R., Stryer, D. B., & Clancy, C. M. (2003). Practical clinical trials: increasing the value of clinical research for decision making in clinical and health policy. *JAMA*, *290*(12), 1624–1632. https://doi.org/10.1001/jama.290.12.1624

Vacha-Haase, T., Tani, C. R., Kogan, L. R., Woodall, R. A., & Thompson, B. (2001). Reliability generalization: exploring reliability variations on MMPI/MMPI-2 validity scale scores. *Assessment*, *8*(4), 391–401. https://doi.org/10.1177/107319110100800404

Vacha-haase, T., & Thompson, B. (2004). How to estimate and interpret various effect sizes. *Journal of Counseling Psychology*, *51*(4), 473–481.

Vacha-Haase, Tammi, Henson, R. K., & Caruso, J. C. (2002). Reliability generalization: Moving toward improved understanding and use of score reliability. *Educational and Psychological Measurement*, *62*(4), 562–569. https://doi.org/10.1177/0013164402062004002

van Elst, L. T., Maier, S., Klöppel, S., Graf, E., Killius, C., Rump, M., … Philipsen, A. (2016). The effect of methylphenidate intake on brain structure in adults with ADHD in a placebo-controlled randomized trial. *Journal of Psychiatry & Neuroscience : JPN*, *41*(6), 422–430. https://doi.org/10.1503/jpn.150320

Van Noorden, R. (2011). Science publishing: The trouble with retractions. *Nature*, *478*(7367), 26–28. https://doi.org/10.1038/478026a

Van Noorden, R. (2014). Online collaboration: Scientists and the social network. *Nature News*, *512*(7513), 126. https://doi.org/10.1038/512126a

Van Norman, G. A. (2016). Drugs and devices: Comparison of European and U.S. approval processes. *JACC: Basic to Translational Science*, *1*(5), 399–412. https://doi.org/10.1016/j.jacbts.2016.06.003

Vogel, S. A. (2009). The politics of plastics: The making and unmaking of bisphenol-a "safety." *American Journal of Public Health*, *99*(Suppl 3), S559–S566. https://doi.org/10.2105/AJPH.2008.159228

Wallach, H. (2006). Topic modeling: Beyond bag of words. In *Proceedings of the 23rd International Conference on Machine Learning* (pp. 977–984).

Wang, Xiaogang, & Grimson, E. (2008). Spatial latent dirichlet allocation. In J. C. Platt, D. Koller, Y. Singer, & S. T. Roweis (Eds.), *Advances in Neural Information Processing Systems 20* (pp. 1577–1584). Curran Associates, Inc. Retrieved from http://papers.nips.cc/paper/3278-spatial-latent-dirichlet-allocation.pdf

Wang, Xuerui, & McCallum, A. (2006). Topics over time: A non-Markov continuous-time model of topical trends (pp. 424–433). Presented at the Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM. https://doi.org/10.1145/1150402.1150450

Wasserstein, R. L., & Lazar, N. A. (2016). The ASA's statement on p-values: Context, process, and purpose. *The American Statistician*, *70*(2), 129–133. https://doi.org/10.1080/00031305.2016.1154108

Weyandt, L. L., Oster, D. R., Marraccini, M. E., Gudmundsdottir, B. G., Munro, B. A., Zavras, B. M., & Kuhar, B. (2014). Pharmacological interventions for adolescents and adults with ADHD: stimulant and nonstimulant medications and misuse of prescription stimulants. *Psychology Research and Behavior Management*, *7*, 223–249. https://doi.org/10.2147/PRBM.S47013

Wicherts, J. M., Bakker, M., & Molenaar, D. (2011). Willingness to share research data is related to the Strength of the Evidence and the quality of reporting of statistical results. *PLOS ONE*, *6*(11), e26828. https://doi.org/10.1371/journal.pone.0026828

Wilcox, D. (2015). Country-of-origin bias: A literature review and prescription for the global world. In *Marketing, Technology and Customer Commitment in the New Economy* (pp. 86–96). Springer, Cham. https://doi.org/10.1007/978-3-319-11779-9_35

Wiles, K. (2014). Why high-profile journals have more retractions. *Nature News*. https://doi.org/10.1038/nature.2014.15951

WMA - The World Medical Association-WMA Declaration of Helsinki – Ethical Principles for Medical Research Involving Human Subjects. (n.d.). Retrieved March 11, 2018, from https://www.wma.net/policies-post/wma-declaration-of-helsinki-ethical-principles-for-medical-research-involving-human-subjects/

Wolfson, M. (2017). *The Fight Against Big Tobacco: The Movement, the State and the Public's Health*. Routledge Press.

Young, S. N. (2009). Bias in the research literature and conflict of interest: an issue for publishers, editors, reviewers and authors, and it is not just about the money. *Journal of Psychiatry & Neuroscience : JPN*, *34*(6), 412–417.

Zhao, H., Liang, F., Fang, Y., & Liu, B. (2017). Application of grading of recommendations assessment, development, and evaluation (GRADE) to the guideline development for clinical practice with acupuncture and moxibustion. *Frontiers of Medicine*, *11*(4), 590–594. https://doi.org/10.1007/s11684-017-0537-4

Zimbardo, P. (2002). American psychologist task force report: Clarifying mission, coverage, communication, and review process. *American Psychologist*, *57*(3), 213–214.

APPENDIX A

ETHICS IN THE APPLIED SCIENCES

Ethics

Ethics is broadly defined as a normative evaluation used to distinguish right from wrong based on varying dimensions, such as: morality, justice, obligation, and self-righteousness (Baier, 1958). Ethics provides a set of rules or guidelines that, when upheld, attempt to protect the fundamental interests of the greater good (Corrigan, n.d.). Many professional fields uphold codes of ethics to guide their professionals (e.g. law, education, medicine, health education, among others) due to the varying emergent issues for which (a) no single, clear-cut answer exists, and (b) some form of moral reasoning regarding context- specific dilemmas is warranted (Benson, 1989).

Codes of ethics are intended to assist decision-making related to ethical dilemmas using rational and logical processes to arrive at the best course of action. Allen (2012) argues three conditions, at least, must be satisfied for a situation to be classified as an ethical dilemma: (1) A person, or agent, must choose a course of action on a subject. If a situation does not require an agent to make a choice, the situation is not a dilemma. (2) There must be varying options for the agent to choose, with numerous potential outcomes. And, perhaps most importantly, (3) regardless of the decision made by the agent, there will, at some level, be some loss. In other words, a perfect solution is not available.

In his book "Moral Reasoning" (1980), Graissing gives several such examples to illustrate the nature of theoretical ethics and the inherent difficulty associated with solving ethical dilemmas. Perhaps one of the most famous examples of a classic dilemma is the Prisoner's Dilemma (Nowak & Sigmund, 1993). In this example, imagine you and a colleague are arrested on unspecified charges, but there is not enough evidence to convict both. You are, then, given two potential options to escape sentencing: (1) remain silent, or (2) betray your colleague. Your actions have one of the outcome conditions: (1) If you betray your colleague and he/she stays silent, you go free and the colleague serves three years in prison, if your colleague betrays you and you stay silent, your colleague goes free and you serve a three-year prison term. (2) If both you and your colleague betray each other, you both serve three years in prison, and (3) If you and your colleague both remain silent, each will only serve one year in prison. Regardless of the decision, there is loss at some level, to either you as the primary agent, or to your colleague.

When presented with an ethical dilemma such as the one above, the agent must rely on a series of cognitive processes to arrive at a *perceived* optimal answer— these processes can be facilitated by invoking professional codes of ethics or personal ethics as guidelines (Wilshere, 1997). The purpose of the Prisoner's Dilemma, for example, is to explore the complexity of cooperative behavior through an example in which, logically, the prisoner's optimal solution would be to betray the other partner (Fehr & Fischbacher, 2003). However, despite appearing to be the optimal solution, there

remains the risk of both partners betrayed each other and both get sent to prison (Molovsky, 2014). There is, therefore, no solution in which all parties benefit equally.

Applied Ethics

While theoretical situations such as the Prisoner's Dilemma are some of the most notably popular examples of exercises in ethical thought, they cannot escape one central concern— the dilemmas and solutions themselves, like much work in classical ethics, are merely conceptual (Son, 2014). Rarely, if ever, do these theoretical ethical dilemmas occur in real world scenarios or applications, posing problems for fields in which its professionals' interactions with people warrant much practice and training.

Professional fields in the applied sciences, such as Health Promotion, Public Health, and Medicine, among others, are ethically driven and stem from classic ethics theory, but they embody inherent needs that cannot be met by theoretical/classic ethics, alone. In other words, the ethical choices being made by agents in applied sciences have real-world ramifications for living populations. Therefore, those fields find themselves resorting to more practical applications of ethical principles. In other words, the formal and informal codes that guide normative decisions in the applied fields are rooted in a popular mid-20th century shift in ethical discourse to what is now commonly known as 'applied ethics.'

Applied ethics is a branch of ethical thought that departs very slightly from traditional Western ethics theory (Rest, 1994). First popularized in the mid-20th century, applied ethics specifically seeks to apply available ethics theory and

frameworks to controversial intra-personal, inter-personal, technological, professional and governmental contexts (Peterson & Ryberg, 2016). Because applied ethics goes beyond classical thought and is centered on decision-making that affects populations, the propositions of applied ethics were eventually incorporated into applied sciences and social sciences (as they deal almost exclusively with human populations), such as Public Health, Medicine, Education, and Health Education, among others.

Bioethics, for example, is one field in applied medical sciences with varying degrees of frameworks, protocols, and debates, aimed at keeping medical research sound and participants in research, or patients, safe from harm. Today the majority of medical policies enacted in the US are inherently following various codes of bioethics (Mbidde, 1998). These codes have been shaped by several advances, dilemmas, and ethical breaches among groups, which forced discourse on ethical rights and considerations (e.g. abortion, stem cell research, euthanasia, among others) (Marshall, 2000). In effect, when a situation warrants ethical discussion within the medical field (due, in part, to concerns over someone's ethical rights), protocols and policies established via bioethical criteria serve as shields to protect certain groups' basic human rights.

Ethics and Human Subjects in Research

Tuskegee Syphilis Study. Perhaps the most notorious and highly studied domestic example of an ethical breach that led to policy development is the Tuskegee Syphilis Study (TSS) and the lasting impact it had on patient rights (Brandt, 1978,

Smith, 1999, Daughtery-Brownrigg, 2012). TSS was a 40-year federally-funded (U.S. Public Health Service) clinical investigation examining the spread and progression of untreated syphilis among rural African American men in Alabama. The participants were deceived into believing their involvement in the investigation would result in free governmental health care and access to medication that was too expensive for that population.

The brazen withholding of information and treatment of participants, once penicillin became available, and the direct and purposeful exposure to harm was viewed as a cruel ethical breach (once the public became aware of the experiment, in 1972; McCallum, Arekere, Green, Katz & Rivers, 2006). The long-term effect of the study eventually led to a congressional court-filed class-action lawsuit settling for 10 million dollars, lifetime medical care, and one of the most important components of future bioethics and research policies: informed consent (Faden & Beauchamp, 1986).

The TSS is an important example, of several, regarding how applied bioethics is used to address a moral dilemma in a real-world situation. In this case, the dilemma was a clear case of deception, withholding of information, and withholding treatment from a curable disease. Unknowing participants were harmed directly by being exposed to untreated syphilis without full disclosure of the intended outcomes of the study. From an ethics standpoint, the moral debate centers on what does one value more: the pursuit of knowledge and science for the sake of science or the promotion of health for the sake of human life? (Katz, 1993).

In effect, the resulting probe, discourse, and debate deemed the harm to participants was so high it moved beyond a threshold of acceptability. Thus, using ethical consideration as a moral code, coalitions of professionals were formed in the 19070's to generate policy to protect future human subjects from experiencing similar harm (Jonsen & Butler, 1975). Today, for example, academic institutions' Institutional Review Board and its oversight of studies, along with informed consent forms used to indicate voluntary participation in research studies, are mainstays in the bioethical process of conducting scientific experiments involving human subjects (Corrigan, 2003).

Abortion. A more modern, yet still highly contentious example of bioethical dilemmas influencing policy in the US is abortion (Jafe, Lindheim & Lee, 1981). Abortion had been a point of contention country=wide since the late 1800's, yet it garnered national spotlight only in the 1960's when it became a felony to terminate a pregnancy without special circumstance, in 49 states and in Washington DC (Petchesky, 1984). After numerous attempts by coalitions to reverse the anti-abortion positions, the US Supreme Court eventually agreed to review the landmark abortion case Rowe vs. Wade in 1971, and ultimately ruled that criminalizing abortion was unconstitutional; women had the right to both choice and privacy when seeking an abortion (Garrow, 2015).

Though the Supreme Court issued a ruling on Row versus Wade, the backlash, debate, and attempts to overturn the decision remain. The resulting aftermath centered,

primarily, on the ethical considerations from two opposing sides: (1) one side claiming it is unethical to terminate a pregnancy, as the fetus is a person from the moment of conception and, thus, has the right to live, and (2) the other, that is an ethical violation to criminalize women's decisions about their own bodies (Garrow, 2015).

As with all ethical dilemmas, it is important to note that arriving at a fair ethical judgement involves a lot of time and consideration into distilling right from wrong. Regardless of whether ethics helped someone arrive at a certain position, others are prone to disagree with that decision. Therefore, if there are moral differences, which, more often than not, there are, a perfect solution is not possible (Hanson, 2014). With abortion, for example, policy recommendation can always be overturned if a new governing body's moral compass differs from the one preceding it.

What TSS and abortion both exemplify is that, within the context of applied sciences, there are serious, sustained ethical concerns — quite different from those explored in classic ethics – affecting human populations. Therefore, when approaching an ethical dilemma in the context of Health-related research there must be consideration for all sides involved in the dilemma before making a judgement on which position an agent morally aligns. Failure to properly think and reason through an ethical dilemma could pose undue harm on certain groups— especially as newer ethical dilemmas continue to push boundaries over what is and is not morally acceptable (e.g. stem-cell research, euthanasia, cloning, among others).

Ethics and Research Methods

While science continues to, and will constantly face ongoing ethical debates regarding human subjects (and animals) who take part in scientific experiments, it is also continually facing new ethical challenges related to the practice of science, itself (Ioannidis, 2005). In other words, the manner in which science is conducting itself in the 21st century is raising eyebrows among scientists and ethicists, above and beyond concerns with the well-being of research study participants (Hubbard, 2015).

Concerns with the practice of science began back in the 1980's when Shweder & Fiske (1984) argued there had to be some level of reform to prevent the spread of unscrupulous science. Three decades later, and little has changed to assuage concerns of an emergent crisis in the social sciences on multiple fronts— such as problematic reporting metrics, fabricating data, among other concerns. More contemporary voices have posited similar concerns about the future of science[38]. Further, editorial boards such as the American Psychological Association (APA) and the Basic and Applied Social Psychology Journal (BASP) have requested modest reforms to prevent a crisis. For example, the APA Special Task for Statistical Inferencing (1999) argued for the inclusion of confidence intervals and effect sizes in research reports to uphold overall credibility. BASP, on the other hand, opted to make significance tests optional if (a) data were made open-access to all interested parties and (b) authors reported effect sizes to substantiate results.

---

[38] E.G. Nuzzo, 2014; Ioannidis, 2005; Epstein 2014; and Hubbard, 2015.

However, such recommendations for editorial reform are just that— recommendations. Researchers have full liberty to abide by or to ignore suggestions made by editorial groups or by other scholar seeking reform to the current system (Resnik, n.d.). Further, researchers also have full academic freedom to construct and execute a study as they see fit. Even if the recommendation is intended to help scientists, and the credibility of science, it is still likely some will continue with practices and protocols and practices that have been used for decades— even if those protocols have been deemed outdated or problematic (e.g. over-reliance on significance testing).

Scientists, however, should aim to produce ethically sound studies across all areas of research— such as patient and animal rights and accurate reporting of information. Today, there are continued concerns within the scientific community— especially regarding methodological decision making and reporting practices— that science, as an enterprise, is facing a growing lack of legitimacy (Hubbard, 2015). Those concerns stem from an emergent ethical dilemma— in which the pressure to produce sharp, cutting edge science prompts many to choose: (a) career? Or (b) rigorous science?

And, as science continues facing poor credibility, there are further ethical implications for the greater public who rely on the production of science to improve quality of life and potentially guide public policy. Therefore, I contend what we, as scientists, are facing is an emergent ethical dilemma regarding methodological

reporting and decision making. One that should be studied further in order to fully understand the unique complexity by methodological reporting and decision making.

APPENDIX B

ASSIGNING A GRADE

The text below is transcribed from an online-published, 2013 University of Exeter Power Point presentation on GRADE and how to assign a GRADE on five key dimensions: (1) Risk of Bias, (2) Inconsistency, (3) Indirectness, (4), Imprecision, and (5) Publication Bias. Since this presentation, the Cochrane Institute has developed a point-and-click interface, GRADEpro, intended to make the process of assigning a GRADE more efficient (Long, 2018).

For each domain, listed below, users of GRADEpro can choose, based on value-judgement, one of the following from a drop-down menu: (1) Not Serious, (2) Serious, or (3) Very Serious. The selection from each domain's drop-down menu will affect the computer-calculated GRADE automatically. In other words, GRADEpro calculates the final GRADE for users.

• STEP 1 Assign an *a-priori* ranking of "high" to randomized controlled trials and "low" to observational studies. Randomized controlled trials are initially assigned a higher grade because they are usually less prone to bias than observational studies— on GRADEpro, users select the study design from a drop-down menu, and the a-priori GRADE appears on screen.

• STEP 2 "Downgrade" or "upgrade" initial ranking. It is common for randomized controlled trials and observational studies to be downgraded because they suffer from

identifiable bias. Also, observational studies can be upgraded when multiple high-quality studies show consistent results Using GRADEpro.

• Reasons to "downgrade"

  • Risk of bias – Lack of clearly randomized allocation sequence – Lack of blinding – Lack of allocation concealment – Failure to adhere to intention-to-treat analysis – Trial is cut short – Large losses to follow-up—external factors, such as funding source and conflicts of interest.

  • Inconsistency: When there is significant and unexplained variability in results from different trials Using GRADEpro.

  • Indirectness of evidence can refer to several things: – An indirect comparison of two drugs. – An indirect comparison of population, outcome or intervention.

  • Imprecision: When wide confidence intervals mar the quality of the data.

  • Publication bias: When studies with "negative" findings remain unpublished Using GRADEpro.

• Reasons to "upgrade"

  • Large effect: When the effect is so large that bias common to observational studies cannot possibly account for the result.

  • Dose-response relationship: When the result is proportional to the degree of exposure.

  • All plausible confounders would have reduced the treatment effect: When all possible confounders would only diminish the observed effect and it is thus likely that the actual effect is larger than the data suggests.

• STEP THREE Input value judgements into GRADEpro. Based on decisions to upgrade or downgrade quality of evidence, GRADEpro will automatically calculate a final GRADE:

Final GRADE ranking		
High	⊕⊕⊕⊕	We are very confident that the effect of the study reflects the actual effect
Moderate	⊕⊕⊕	We are quite confident that the effect in the study is close to the true effect, but it is also possible it is substantially different
Low	⊕⊕	The true effect may differ significantly from the estimate
Very low	⊕	The true effect is likely to be substantially different from the estimated effect

Long, L. (n.d.). Understanding GRADE: An Introduction. Power Point. Retrieved February 15, 2018, from https://medicine.exeter.ac.uk/media/universityofexeter/medicalschool/events/docs/Oct_2013_GRADE.pdf