4-19-2022 2:30 PM

Spin and Distortion in Surgical Trials

Andrea Mataruga, *The University of Western Ontario*

Supervisor: Martin, Janet, *The University of Western Ontario*

A thesis submitted in partial fulfillment of the requirements for the Master of Science degree in Epidemiology and Biostatistics

Abstract

Research problem: Randomized controlled trials (RCTs) are essential; however, their validity can be threatened through distortion or spin. This study quantifies publication bias and distorted outcome reporting.

Methodology: All surgical RCTs registered on ClinicalTrials.gov from 1997-2017 were identified and a sample was obtained through random and intentional selection. Failure to publish (proportion of studies that remain unpublished), outcome distortion (changing intended outcomes), and spin (distorted presentation) were explored. Comparisons were made for positive versus negative studies and for high-income (HICs) versus low-middle income countries (LMICs).

Results: In total, 13,761 RCTs were registered (median enrollment size = 96, 94% from HICS). From a sample of 5,094 studies, 1,718 of them were published (34%). In total, 62% of published conclusions declared a significant difference (1,058/1,718), of which 41% (436/1,058) had “turned” positive due to spin or distortion.

Conclusion: While a large volume of RCTs have been registered, many remain unpublished. High proportions of spin and distortion raises concerns for validity of the evidence base.

Keywords

surgery, randomized controlled trials, research integrity, publication bias, spin

Summary for Lay Audience

Randomized controlled trials (RCTs) are a necessary tool to support evidence-based decision making in clinical practice. To provide valid information, it is essential that RCTs publish all their results and report the outcomes without bias. In this study, we examine characteristics and trends of surgical RCTs including evidence of publication rates and outcome distortion. To do this, we retrieved all surgical RCTs registered on ClinicalTrials.gov between 1997 and 2017. We examined the research base describing the global spread of studies and volume of study registration across the years. After this, we took a sample of the studies to search for publication and compared the outcomes that were intended to be reported with the outcomes the publication reported. Differences between intended outcomes and reported outcomes are known as outcome distortion, and misrepresentation of results in the presentation of the outcomes is known as spin. These characteristics were compared between positive and negative studies, as well as for high-income countries (HIC) and low-middle income countries (LMIC). In total, we retrieved 13,761 RCTs that met the inclusion criteria. These studies had a median enrollment size of 96 that decreased across the years, and mainly pertained to HICs (94%). The number of studies registered each year was increasing; however, the proportion labelled "Completed" was decreasing. The sample of studies consisted of 5,094, where 1,718 were published (34%). Of published studies, conclusions with significant differences were declared in 62% of the studies (1,058/1,718), of which 41% (436/1,058) turned positive due to spin or distortions of outcomes. HIC had a higher failure to publish rate than LMIC (63% vs. 49% unpublished), but overall had more studies published, and increasing proportion of positive studies published. The high proportion of studies in HIC is a concern for global generalization and should be more reflective of where majority of the world's population lies. Additional concerns for research integrity and validity of the evidence base lie within the large proportion of unpublished studies (66%) despite the large volume of registered studies. In order to ensure adequate decision making, further efforts to ensure publications are conducted without biases are necessary.

Co-Authorship Statement

The conception, design, and execution of this study was performed by Andrea Mataruga, Dr. Janet Martin, and Dr. Eunice Chan. Dr. Neil Klar was a thesis committee supervisor and provided comprehensive feedback. All authors contributed to interpreting data, editing the thesis, and giving final approval for submission.

Acknowledgments

This thesis would not have been possible without the endless care and support from my friends, family, and team. Their kindness and generosity have allowed me to maintain my happiness and stay motivated throughout this entire process, especially considering the difficulties brought forth by the pandemic for us all.

I must express my sincere gratitude to my supervisor, Dr. Janet Martin, for the exemplary mentorship, and continued support over these last 2 years. Thanks to her knowledge, encouragement, and guidance, I was given the strength and opportunity to create globally impactful research that never would have been completed otherwise. I have learned so many lessons throughout this process that I will carry with me through the rest of my career.

I would also like to thank my advisory committee member Dr. Neil Klar for his support and insightful feedback to further improve this work. I must also thank Dr. Eunice Chan who I am indebted to for continuously providing support and guidance throughout this entire process, and at all hours of the night. Your patience and expertise have kept me motivated and allowed me to bring forth the best version of this project. I must also give thanks to my team who has been dedicated to providing their generous support including Jessica Moodie and the group of volunteers at MEDICI.

A big thank you to my friends and colleagues at Western University for an unforgettable experience. The light-hearted chats, thoughtful advice and words of encouragements have been an integral part of my success. Lastly, a big thank you to my friends, partner, and my family, Dusan, Sandra, Marina, and Ivana Mataruga, for their continued love and support.

Table of Contents

List of Tables

List of Figures

List of Appendices

List of Abbreviations and Acronyms

RCT	Randomized Controlled Trial
HIC	High-income country
LMIC	Low-middle income country
HDI	Human Development Index
WHO	World Health Organization
ICMJE	International Committee of Medical Journal Editors
ISRCTN	International Standard Randomised Controlled Trial Number
FDAAA	Food and Drug Administration Amendments
CI	Confidence interval
IQR	Interquartile range
JIF	Journal Impact Factor
NCT	National Clinical Trial
WoS	Web of Science
NA	Not applicable
SE	Standard error
JAMA	Journal of the American Medical Association
UN	The United Nations
SAR	Special Administrative Regions
ENT	Ear, nose, throat
GI	Gastrointestinal
HCC	Hepatocellular carcinoma
PMID	PubMed Identifier
PMCID	PubMed Central Identified
URL	Uniform Resource Locators
POI	Primary outcome intended
POR	Primary outcome reported
PO	Primary outcome

Chapter 1

1 Literature Review & Introduction

1.1 Surgery

Derived from the Latin and Greek words for “hand work”, the term surgery is defined as the treatment of injuries or diseases in people or animals by cutting open the body and removing or repairing the damaged part.[1,2] Though definitions vary across dictionaries, the commonalities are that surgery is a procedure to cure or treat the burden of disease using techniques related to cutting the body open.[1] This procedure is relevant for any and all body parts such as Gastroenterology, Ophthalmology, Otolaryngology and Cardiology to name a few. The term “surgery” can also be referred to as an operation, a surgical procedure, and other specific terms more specific to the area of the body such as an appendectomy referring to the removal of the appendix, or a cesarean section referring to childbirth through an incision of the abdomen. While the field of surgery itself has evolved extensively, the role of the surgeon has adjacently evolved, moving from a largely technical role focused on performing the operation, to a position of both technician and doctor managing the diseases within and beyond the operating room.[3]

Throughout history, advancements in technology and medicine have driven the evolution of surgery, predominantly as a result of the technology boom of the last 200 years.[1] Although humans have been performing surgeries for centuries with the first known record of surgery to be in 600 B.C, it wasn’t until the “Industrial age” of the mid-1800s and the introduction of new instrumentation and anesthesia that the field of surgery really started to develop.[1,2] Following this age, technology and medicine continued to grow coming into the “Information Age” of the 1950s.[1] This allowed for more invasive open techniques as a result of the discovery of antibiotics and the utilization of intravenous fluids, and eventually video cameras, robotic systems, and thus, minimally invasive procedures that revolutionized surgery as a whole.[1]

1.1.1 Access to Safe Surgery

According to a World Health Organization (WHO) modelling study performed in 2008, there are an estimated 234.2 (95% CI 187.2–281.2) million major surgeries performed each year worldwide.[4] Of this 234.2 million, a disproportionate 73.6% were estimated to occur in high or middle-income countries, while only 30.2% of the world's population resides in these locations.[4] The risk of death and complications associated with surgery has lessened through time; however, there is still an estimated permanent disability or death rate of 0.5-0.8% after major surgery in high-income countries, and the risk remains several-fold higher in low-middle income countries for procedure-matched and risk-adjusted comparisons.[4] Surgical complications occur in up to 25% of patients, with at least half of them considered to be avertable.[5] Simultaneously, there is a close correlation (R^2=0.996) between per-capita expenditure on health and the volume of surgery.[4] It is now estimated that 16.9 million lives are lost every year due a lack of access to surgical care, indicating an urgent need for proportionate and increased surgical care available, particularly in resource-restricted countries.[6]

1.2 Randomized Controlled Trials

Randomized controlled trials (RCTs) are a type of experimental study, where participants are prospectively randomized to either intervention or control group, to (theoretically) ensure groups are comparable for all aspects other than the difference provided by the intervention under study.[7] The RCT is often referred to as the gold standard for ascertaining efficacy and safety of an intervention, allowing for reduced confounding and improved measurement of intervention effectiveness while reducing bias through the randomization process.[7] In the hierarchy of evidence to inform decisions about what works, well-conducted randomized trials (or systematic reviews of all relevant randomized trials) remain at the top of the hierarchy in terms of evidence validity.[8] Randomized allocation is often done using automated randomization tools such as computerized random number generation to ensure allocation is beyond the influence of investigator. In addition, allocation of the participants is ideally blinded to the

participants, researchers, and the medical professionals applying the interventions in order to minimize threats to randomized allocation.[7] However, in the field of surgery, blinding of patients, clinicians and investigators is particularly challenging since surgery is a procedure which is difficult to emulate with a "placebo" or matching "sham" control in a manner that reliably hides knowledge of which intervention was received. If the trial addresses a surgical intervention compared to non-surgery, it can be challenging to conceal if a surgery has taken place due to the invasive nature of them, and ethical concerns may arise in proposing sham surgeries. Additionally, despite being the gold standard of clinical research, RCTs may have other drawbacks including high resource demands, and threats to validity if there is attrition during longer term follow up. Furthermore, barriers to recruitment may result in a population that is less generalizable.[8] As the highest level of evidence, performing RCTs are the ideal for all areas of research and undoubtedly in surgical research due to the invasive nature and high risk of harm, warranting a need for high quality evidence-based research.

A total of 386,745 studies from 219 different countries have been registered between 1997 to the present day (August 15, 2021), and approximately half of these are randomized trials.[9] The number of studies registered has increased over time, possibly due to increased mandates for registering. In 1997, there were only 323 registered randomized trials, whereas in 2020 there were 14,313. Registrations in 2021 will soon exceed those in 2020 since there have already been 13,025 by mid-August 2021.

1.3 Trial Registration

As of September 2005, The International Committee of Medical Journal Editors (ICMJE) has defined a list of 6 web-based registries without for-profit affiliations considered acceptable for trial registration : www.anzctr.org.au, www.clinicaltrials.gov, www.ISRCTN.org, www.umin.ac.jp/ctr/index/htm, www.trialregister.nl and https://eudract.ema.europa.eu/.[10] Additionally, any registries within the WHO International Clinical Trials Portal will also be accepted.[10] For our research, we selected ClinicalTrials.gov since it is has been the most commonly used registry over the past 20

years, spanning 219 countries worldwide and receiving about 4.5 million visitors each month.[8] It is a well-established resource, featuring both publicly and privately funded projects, that is intended for providing a public platform for patients, researchers, health care professionals, and the general public to view registered, ongoing, and completed clinical trials.[11] Due to the public nature of the site, researchers can be held accountable for maintaining the intentions of their trial consistent, and transparently reporting any changes to the original plan. Trial registration is intended as one safeguard to up-hold transparency and integrity of research, and theoretically should aid in the reduction of the net adverse effects of reporting bias, publication bias, and spin on the evidence base. Trial registration also contributes to other benefits including redundant duplication of clinical trials and unnecessary research waste. A further benefit may include opportunity for patients and investigators to publicly access recruiting studies to see if they qualify for enrollment. Lastly, clinical trial registries may facilitate ethics review boards and granting agencies to make decisions about new proposed clinical trials.[10]

According to the Food and Drug Administration Amendments Act (FDAAA) that governs ClinicalTrials.gov, all studies that are controlled clinical investigations of any drug, biological product, or medical device, excluding feasibility trials, and are initiated either after September 27, 2007, or prior to and still ongoing past December 26, 2007, must be registered (exceptions and further details can be reviewed within the FDAAA 801 document).[12] The ICMJE follows similar guidelines for trials to be eligible for publication in one of their journals.[10] They state that trials beginning on September 13, 2005 or later for any intervention will be considered for publication in an ICMJE journal only if they are registered in one of the accepted registries and contain all the information within their registration (exceptions and further details can be reviewed within the ICMJE information pages).[10] A typical trial registration will define participant enrolment criteria, condition or disease, intervention and comparators, primary and secondary outcomes, study locations along with contact information, and other relevant information.[11]

1.4 Research Integrity Within Publications

1.4.1 Pressure to Publish

Within the academic community, there is an immense pressure to produce a high volume of publications, or highly cited publications as these are rewarded by current evaluation frameworks for promotion and tenure.[13] Research publication remains one of the most highly valued measures of academic success, potentially lead to increased funding for the institution and an overall increased status of the institution on the global stage.[13] This incentivized structure has contributed to the increase in research publications over the years. A 2014 study from Cornell University examining global volume of scientific publications on the Web of Science (WoS) database showed an exponential growth between 1980 to 2012.[14] They claim that in the mid-1700s the growth rate was less than 1%, increasing to 2-3% by the mid-1900s, and then increasing to 8-9% by 2012.[14] Though the study also claims that these values are likely incomplete, potentially overestimating the growth rate due to the difficulty of tracking studies from the earlier years, there is clearly an exponential growth of published scientific literature in recent decades.[14]

The pressure to publish has been a contributor to unethical research practices, breaching research integrity, and increasing overall wasteful research.[13] It has been stated that only 42% successfully published articles receive more than 1 citation, some of which are self-citations from the authors or journals (5-25%).[13] Additionally, less than half (45%) of the articles published in the top 4500 scientific journals receive a single citation within the first 5 years of publication.[13] This suggests that much of the published work goes unnoticed, and may fail to meaningfully contribute to the chain of knowledge translation and evidence-informed decision making.[13]

1.4.2 Spin

For patients to receive the best possible care, it is vital that the research pool is continuously updated, and the best practices are in use. With the pressure to publish,

some researchers may be incentivized to adopt questionable or unethical practices to increase likelihood of publication.[13] Questionable or unethical practices may range from more nuanced practices such as selective reporting of positive or "more interesting" results, switching non-significant originally planned primary outcomes with significant secondary outcomes, failing to publish study results, over-interpreting results to align with preferred conclusions, indiscriminate statistical analysis to 'find' significant outcomes, non-transparent declaration of statistical plan or changes to original protocol, all the way to potential extreme of overtly nefarious practices such as falsification of results, fraudulent data manipulation. In effect, anything that researchers do either subconsciously or consciously to collect or interpret data in ways that misrepresent the original intentions of the research, or selective reporting favourable results, leads to questionable research outcomes.[13]

This misrepresentation of research is commonly known as "spin", and involves the deceptive use of presentation or language to display results that emphasize the benefits of favourable results or mask unfavourable results, despite a non-statistically significant outcome.[15] Spin tactics can be used to either alter the outcome completely, or simply mislead the reader with the choice of wording or display of the results.[15] A study by Boutron and colleagues analyzed the extent of spin in published RCTs with non-significant primary outcomes, where spin was defined as: a "specific reporting strategy" involving result misrepresentation that can "highlight that the experimental treatment is beneficial, despite a statistically nonsignificant difference for the primary outcome, or to distract the reader from statistically nonsignificant results".[15] They identified spin within results and conclusions sections in 27% of published study abstracts, and in 42% of the main manuscripts.[15] The act of spinning research results can be either conscious or subconscious; however, regardless of the specific intention or motive, the far-reaching health and social consequences of spin on interpreting the evidence base should not be minimized. It is essential that researcher education, awareness, and skills be increased to recognize and mitigate spin. Additionally, meta-researchers should evaluate the current evidence base to quantify spin more transparently, and highlight inconsistencies between the planned research, reported, results, and interpretation of findings.[15]

Spin can be present in many ways. Commonly, spin is incurred through misleading language such as using causal language to infer beyond what the results have shown. Another common approach is for spin to be incurred through distraction from non-significant or non-favourable results, often with a focus on other secondary statistically significant results cherry-picked from a long list of potential outcomes that increase risk of false positives with indiscriminate use of analytic flexibility, also known as 'researcher degrees of freedom' and the 'garden of forking paths'.[16] A number of techniques used by researchers may be used consciously or subconsciously to deflect attention from non-significant results to 'turn' the story toward statistically significant results, including focusing on secondary outcomes or subgroup analyses, interpreting statistically non-significant results in a manner suggesting at equivalence between the interventions, or eliminating or downplaying adverse events while placing emphasis on the positive or beneficial results.[16] Spin can ultimately be present in any section of the research publication, from the title to the conclusion; however, it is predominantly present in places where there is an opportunity to interpret results, such as in the title, discussion, and the conclusion sections.[16] Research that is published or presented with spin is an inaccurate representation of the results, and leads to detrimental impacts on health and society through misguided decisions and harm to patients and society at large.[16]

1.4.3 Outcome Distortion

Outcome distortion is related to the concept of spin; however, instead of misrepresentation of outcomes through highlighting significant results or distracting the reader from non-significant results often through the use of language, outcome distortion refers to the selective reporting of numerical outcomes through swapping non-significant outcomes with significant ones or removing non-significant outcomes as a whole.[15,17] Generally this is seen as the omission of non-significant outcomes referred to as selective reporting or switching a primary outcome that is non-significant with a secondary outcome that is significant to present positive results though these were not the originally

intended outcomes.[17] For example, this can include the omission of a negative result, but it can also include the omission of an analysis such as an adjusted analysis or intention-to-treat if it does not agree with the rest of the results. In this case, a positive result is one where a statistical difference is seen between the intervention groups, where a negative result is where a null result is observed usually without statistical significance.[18] While positive results may seem interesting or are said to be favourable for publication, non-significant results are equally as important for the evidence base. A 2010 study reported that when comparing published medical studies to their protocols, about 40-60% of them had introduced, omitted or the changed primary outcomes from the per-protocol intentions.[19] Outcome distortion in research can lead to overestimating the positive results, lack of transparency for all the available evidence, and leads to concerns of validity of clinical trials.

1.4.4 Publication Bias and Failure to Publish

There is a common perception, and some analyses have shown, that studies with positive findings have a greater likelihood of being published in a higher-impact journal, and at quicker rates.[18] Ideally, published results will be transparently reported as intended per-protocol despite the directionality of results. However, in a number of cases, spin and outcome distortion are used by investigators to create positive results for publication.[18] A 2014 study provides an overview of empirical evidence for this belief, including "time-lag bias", where negative studies are published with a greater delay, and the "proteus phenomenon" suggesting that delayed negative publications contradict the positive findings creating extreme contradictions in the evidence base.[17] Another 2013 study examining publication rate for drug related clinical trials in Spain found a publication rate of 84.9% (180/212) for studies that concluded positive results, and 68.9% (128/186) for negatively concluded studies. The study also reported that positive studies had a median time to publication of 2.09 years compared to 3.21 years, suggesting a potential impact of positive studies. Negative trials will often be disregarded or never submitted to journals, which leads to the "file drawer problem" of failing to publish. Together, these biases lead to overrepresentation of positive results in the evidence base.[18] With only a subset of the

evidence available to the public, formulation of false conclusions and distortion of meta-analyses may ensue, resulting in unnecessarily repeated future work and wasted research opportunities and resources.[20]

1.5 Rationale

Randomized controlled trials (RCTs) are an integral aspect of evidence-based medicine and are especially important to guide decision making in high stakes areas such as surgical care. However, without research integrity and full transparency within the literature, inadequately performed and reported RCTs can jeopardize the validity of the evidence and threaten decision making. Similarly, it is necessary for research to be relevant and generalizable to the global population where it will be used. This integrated thesis consists of 5 chapters, the first chapter providing background information and rationale for this topic, the three middle chapters each building on the prior chapter to characterize the state of the evidence in the field of surgery, with a focus on degree of outcome distortion and spin in registered RCTs over the past two decades, and the final chapter providing an integrated discussion and future directions.

Chapter 2 will describe characteristics of the entire volume of surgical RCTs registered on ClinicalTrials.gov between 1997-2017, with a focus on the distribution of research production globally, rate of study registration and completion over the years, enrolment size, surgical category, and other basic characteristics. This will provide us with a valuable understanding of the types of procedures of focus, sample size, and where majority of the studies are being produced, and if this is proportionate with where most of the world's population resides. Additionally, we will be able to view trends of study registration over time, and if completion rate is following the same trend. This knowledge will uncover where the gaps in the literature are, and where we need to direct future research.

To analyze research integrity, Chapter 3 will obtain a sample of the full cohort of studies identified in Chapter 2. Using this sample, we will search for published studies to

compare intentions set by the registries to the outcomes that were reported in the studies. Searching for publications will provide us information on the publication rate of studies and identify the proportion of studies that failed to publish. Within the published studies, we will be able to compare the proportion of studies that achieved a positive overall result against a negative overall result and analyze which of these studies kept their intentions from the registries and which turned positive, potentially in the search for an interesting conclusion. Retrieving the proportion of positive studies and the studies that turned positive will reveal the level of distortion, spin and ultimately research integrity held by the evidence base and offer areas where improvement in research reporting may be necessary.

Using the same sample, Chapter 4 will address the same areas for research integrity to explore differences between high-income settings (HICs) versus low-middle income settings (LMICs). Study characteristics, outcome distortion, and spin will be compared for HIC and LMICs. Chapter 2 identified distributions of study registration globally, and now in Chapter 4, we will determine whether rates of publication, outcome distortion, and spin are related to income level of country in which the study was conducted. We will also compare basic characteristics between these groups to see if there is a difference in the studies produced in differing income levels. The aspirational global goal in research should be to produce studies proportionate to where the world's population resides, relevant to burden of disease regardless of income level of the country, so we hope to determine if this is the case.

Together, these chapters will address the questions: What is the state of the evidence base in the field of surgical RCTs? To what degree is registered research completed and published? And of the published research, to what degree is distortion and spin detectable in the results? Altogether, this integrated research will help to inform the integrity of the current evidence base in surgery, and will provide insight for researchers, physicians, and policymakers for future improvement.

Chapter 2

2 Characterizing the Global Body of Registered Surgical Randomized Controlled Trials from 1997 to 2017

A large volume of surgical randomized controlled trials (RCTs) has been registered in ClinicalTrials.gov over the several years, with an increasing amount registered each year. Understanding the characteristics and distributions of the global surgical research is valuable knowledge to understand the gaps that may lie regarding generalizability or applicability of the research, and where further research should focus. The objective of this chapter was to characterize the registered surgical research and identify gaps or disproportions that can be described on a global scale.

2.1 Introduction

Many barriers exist to performing a high-quality study at low risk of bias in the field of surgery. The high level of risk involved with surgery demands a high burden of proof for safety and efficacy, yet surgical trials are particularly difficult to perform since surgical procedures are more involved with added challenges to randomization, blinding and informed consent compared to studies of more simplistic interventions such as drug therapies.[21, 22] To understand the global state of the evidence base related to surgery, it is necessary to evaluate the volume of this research on a global scale to uncover where the inequalities lie and what the trends are over time. More specifically, there is a disproportionate representation of high-income to low-income countries producing the research where tackling these gaps in information will lend to a more equal distribution of research production and in turn limit the bias across economic settings. With this, we will obtain an understanding of characteristics of the studies, changes over time, and thus will have a valuable perspective for the integrity of the research conducted. To our understanding, there has not been a review for the entirety of the surgical research base spanning over 2 decades thus able to analyze the trends and potential changes in the characteristics of the studies over the years. With this said, we aim to describe and characterize the global research trends on human surgical procedures that have been

registered on ClinicalTrials.gov from 1997 to 2017 and identify any gaps or disproportionate representation within the research base.

2.2 Methodology

2.2.1 Search Strategy

A search strategy was formulated to capture all randomized studies that involved a surgical procedure on humans registered on ClinicalTrials.gov database between January 1, 1997, to December 31, 2017. Established in 1997, ClinicalTrials.gov is a database of registered clinical trials for both privately and federally funded studies from 219 countries.[3,4] Initially, all studies up to the end of 2020 were captured in order to identify an appropriate end date. After finding that the median completion time for studies reported on the registry was just over 2 years (25 months), it was estimated that allowing 3+ years for study completion was considered sufficient to capture an accurate and large volume of studies. Consequently, we chose end of 2017 as the end date of our study to allow for sufficient time elapse between trial registration, trial completion, and publication.

The search was conducted in November 2020 using the Expert Search feature in ClinicalTrials.gov and included an exhaustive list of terms describing variations of "surgery", as well as the different types of specific procedures (i.e., caesarean, appendectomy, rhinoplasty, etc.). To ensure that we captured studies that had randomized allocations, the list of search terms was joined with iterations of the term "randomized" using the Boolean operator "AND". The surgical terms were searched within the title, the brief summary, and the designated intervention section of the registry, while the randomized terms were searched within the title and the design allocation sections. Additional restrictions were placed to limit the study type to "interventional" in order to help identify randomized studies, and for dates to fall within the designated date ranges. The exact search strategy used for the search is included in Appendix 1.

2.2.2 Screening for Inclusion

After extracting the studies from ClinicalTrials.gov, they were screened according to a predefined inclusion criterion: the studies must be a randomized allocation design with human patients undergoing a surgical procedure. A surgical procedure is defined as an incision made using a cutting tool while undergoing either local or general anesthesia. The screening was done by one reviewer who examined the pre-extracted pieces of information from ClinicalTrials.gov including the title, interventions, and the arms of the interventions. If it was still unclear whether the trial met the inclusion criterion from these characteristics, the registry was further examined using all other available information. Screening was conducted between November 2020 and January 2021.

2.2.3 Data Extraction

Study characteristics were extracted from the included registered studies, including enrollment size, date of registration, date of study completion, country where the study was performed, status of the study (completed, recruiting, not yet recruiting, enrolling by invitation, active not recruiting, suspended, terminated, withdrawn, and unknown), intervention type, and outcome measures. It was assumed that the registries were updated by the authors and correctly inputted, so we directly used the registry for information unless it was unavailable. While most of the registered studies provided the necessary characteristics, some did not have this explicitly listed in the proper section of the registry. Therefore, a custom Python program that matched key words of the data extracted characteristics to sections within the registry (Appendix 5) was used to retrieve information from other sections of the registry. In addition, published articles were also sought to inform missing fields when possible. We ended the data extraction process in April 2021.

2.2.3.1 Completion Status and Time to Completion

Completion status was defined as “Completed”, “Not Completed”, or “Ongoing” based on the following categories given in ClinicalTrials.gov: “completed” consisted of studies with a ‘completed status’; “not completed” consisted of ‘suspended’,

‘terminated’, ‘withdrawn’, or ‘unknown’ (which includes trials not updated past the recruitment phase, and the status has not been verified within 2 years and the estimated completion date has passed); and “ongoing” consisted of registered studies in the recruitment phase including ‘not yet recruiting’, ‘recruiting’, ‘enrolling by invitation’, ‘active’, or ‘not recruiting’. For the completed studies, time to completion was also calculated by taking the difference between the start and end date published in the registry.

2.2.3.2 Human Development Index and Geographic Region

The country of study conducts, and the respective Human Development Index (HDI) was recorded. In the case that the country was not available from the registry, we attempted to look for the country in the study publication if the publication was available. These countries were then grouped according to the HDI and geographical region. The HDI was obtained from the United Nations 2019 report.[23] If a study involved multiple countries, the average of the countries’ HDI was taken. Countries that did not have HDI available (i.e., countries that no longer exist) were matched with the country that they now belonged with or an average of analogous countries. Along with the HDI, the studies were placed into the universally accepted categories of very high, high, medium, or low HDI.[23] In addition, we categorized the countries where the trials were conducted according to geographic regions designated by the World Health Organization (WHO).[24] These regions include the African Region, the Region of the Americas, the South-East Asian Region, the European Region, the Eastern Mediterranean Region, and the Western Pacific Region. Countries not listed as a part of the WHO regions were placed into categories based on their geographic location.

2.2.3.3 Surgical Categories

For each registered trial, type of surgical procedure performed was extracted from ClinicalTrials.gov according to the following surgical categories: Breast; Orthopedic; Neurology; Transplant; Obstetrics and Gynecology; Plastic; Urology; Pediatric; Cardiac,

Vascular and Thoracic; Otolaryngology; Thyroid; Dental, Oral and Maxillofacial; Colon and Rectal Surgery; Ophthalmic Surgery; Gastroenterology; and General.

2.2.4 Statistical Analysis

Summary statistics (frequencies and proportions) were calculated using Microsoft® Excel (v.16.51), including number of studies reaching completion, the proportion of completed studies, and the average sample sizes for all registered surgical clinical trials by creating various charts using Microsoft® Excel. To visualize the number of studies registered and the relative HDI categories for the countries of registration, a world map was created in Tableau Desktop v.2021.1.5 (20211.21.0819.1914), wherein each country is represented by a bubble, where size of the bubble represents the number of studies registered there, and colour of the bubble represents the HDI level.

2.3 Results

2.3.1 Search Results

Between the beginning of 1997 and end of 2017, a total of 264,301 studies were registered in ClinicalTrials.gov. Of these, 209,209 (79.2%) were labelled as "interventional", and limiting it further, 138,593 (52.4%) were randomized. Searching by our list of terms (Appendix 1) yielded 24,740 (9.4%) surgical RCTs to be screened for inclusion. Of these, 13,761 RCTs met the inclusion criteria, representing 5.2% of all studies registered between 1997 and 2017. Baseline characteristics of the trials are shown in Table 1.

Table 1: Characteristics of Registered Surgical RCTs (1997-2017)

Characteristic	All studies (n=13,761)
	No. (%)
Location by Region	
Americas	5,931 (34.3)
Europe	7,568 (43.8)

Western Pacific	2,545 (14.7)
Eastern Mediterranean	753 (4.4)
South-East Asia	362 (2.1)
Africa	128 (0.7)
HDI[1] Category	
Very High	10,567 (76.8)
High	2,378 (17.3)
Medium	184 (1.3)
Low	35 (0.3)
Surgical Category	
Breast	380 (2.8)
Orthopedic	2,745 (20.0)
Neurology	308 (2.2)
Transplant	738 (5.4)
Obstetrics and Gynecology	1,318 (9.6)
Plastic	256 (1.9)
Urology	713 (5.2)
Pediatric	482 (3.5)
Cardiac, Vascular and Thoracic	2,198 (16.0)
Otolaryngology	312 (2.3)
Thyroid	121 (0.9)
Dental, Oral and Maxillofacial	443 (3.2)
Colon and Rectal Surgery	754 (5.5)
Ophthalmic Surgery	582 (4.2)
Gastroenterology	1,939 (14.1)
General	471 (3.4)
Completion Status	

Completed	7,849 (57.0)
Not Completed	4,307 (31.3)
Ongoing	1,605 (11.7)
Enrollment Size, median (IQR)	96 (52–195)
Enrollment Size, mean (range)	256(1-19,000)
Time to completion (months), median (IQR)	25 (14–45)
Time to completion (months), mean (range)	34 (0-246)

1. HDI = Human Development Index

2.3.2 Human Development Index and Geographic Region

The majority of surgical RCTs were performed in countries with a very high or high Human Development Index (HDI). Together, studies that took place in very high (76.8%) and high (17.3%) HDI countries made up 94.1% of the total number of studies that met the inclusion criteria, leaving studies from countries with medium HDI with 1.3% and low HDI with 0.3% (Table 1). When observing the studies by year, we notice that the first medium-HDI study does not appear until 2001, and the first low-HDI study does not appear until 2007, but the number of studies from both HDI categories has slowly increased over the subsequent years until 2017 (our study end date) (Table 2). A table of the number of studies, HDI value, HDI category, and geographic region for each country can be found in Appendix 3.

For geographic regions, the European Region had the greatest percentage of studies with 43.8%, followed by the Region of the Americas with 34.3%. Very little representation was seen from the Africa Region (0.7%), the South-East Asian Region (2.1%) and the East Mediterranean Region (4.4%) (Table 1). At a country level, the largest proportion of studies came from the United States taking part in 23.3% of the studies.

Table 2: HDI Categories and Enrollment Sizes Distributed by Year

Year	HDI Categories (No.)				Enrollment Size (mean)
	Very high	**High**	**Medium**	**Low**	
1997	27	2	0	0	286.9
1998	50	0	0	0	467.4
1999	46	6	0	0	541.4
2000	61	6	0	0	354.8
2001	110	2	1	0	346.3
2002	139	10	0	0	327.8
2003	186	10	3	0	304.8
2004	260	18	3	0	320.4
2005	341	30	5	0	197.9
2006	400	39	4	0	274.0
2007	482	55	9	1	235.8
2008	616	99	4	2	211.6
2009	678	87	16	2	156.4
2010	724	126	8	2	158.4
2011	858	156	16	7	150.9
2012	806	140	6	4	174.8
2013	886	189	15	3	175.4
2014	928	275	11	2	164.4
2015	964	294	18	4	182.1
2016	977	401	36	4	160.2
2017	1,060	440	29	4	187.1

2.3.3 Completion Status

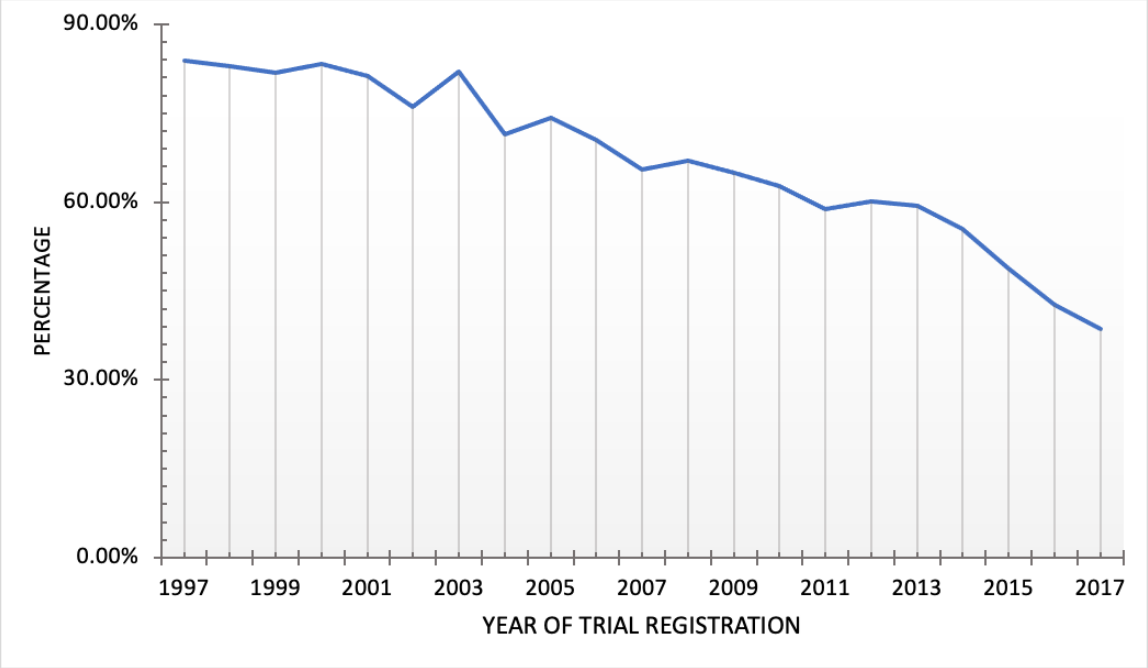


Figure 1: Percentage of studies completed by year of study registration

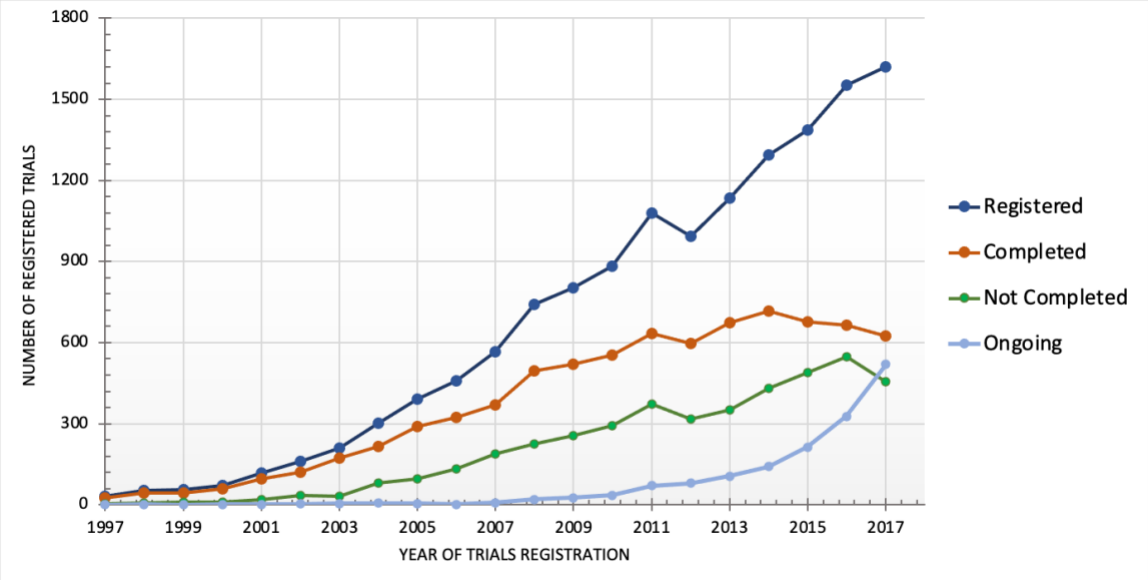


Figure 2: Number of studies according to completion status by year of study registration

The number of surgical RCTs registered each year increases as the years progress (Figure 1), while the proportion completed decreases notably over the years. In total, 57% of registered surgical RCTs were labelled “Completed”. In the inaugural year of the

database, most of the studies that were registered on ClinicalTrials.gov were labelled as "Completed" (83.9%), whereas the percent of completed studies drop by more than half in 2017 with only 39.0% of the studies completed (Figure 2). As expected, the number of ongoing trials increases in more recent years (Figure 1). Time to completion averaged 34 months and had a median of 25 months. The higher mean time to completion compared to the median suggests at a positive skew where more studies will fall to the lower end of the time to completion.

2.3.4 Enrollment Size

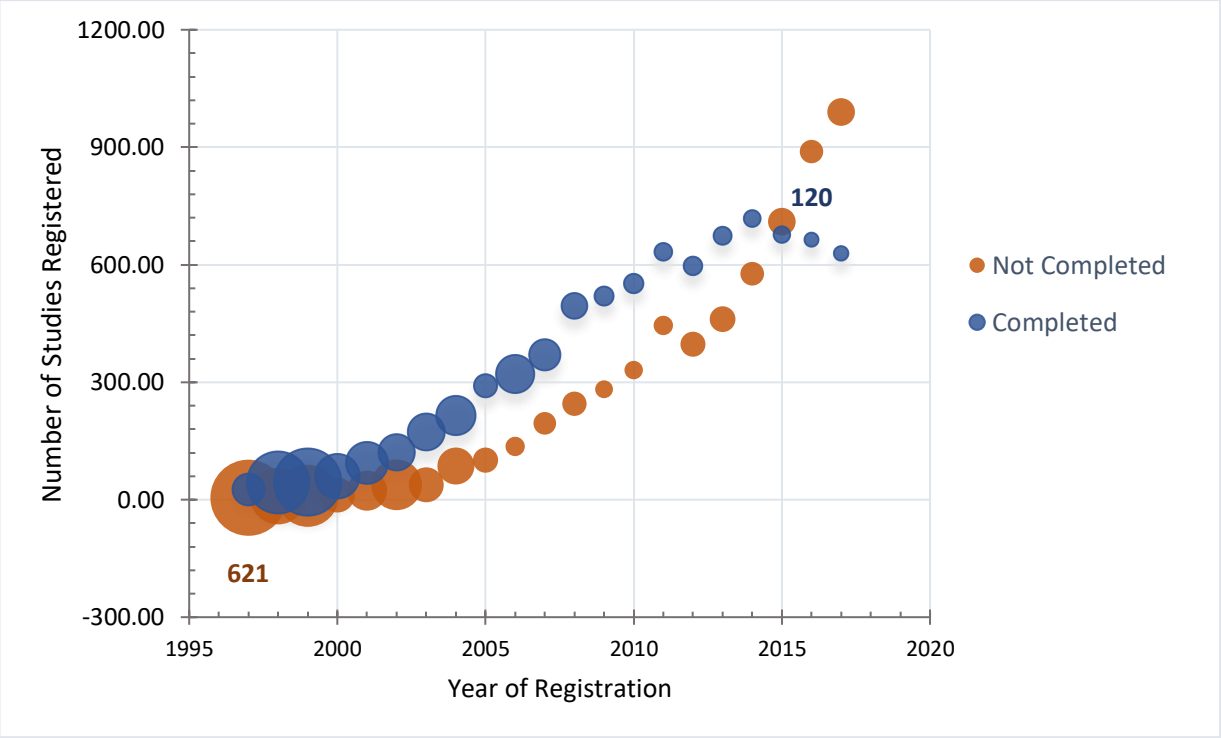


Figure 3: Average enrollment sizes indicated by the size of the bubbles and number of registered studies from 1997 to 2017

The median enrollment size across all studies was 96 (IQR 52–195), and the mean was 256 (Table 2). As seen by the decreasing size of the bubbles in the later years, enrollment size has decreased over time while number of registered studies has increased (Figure 3). The overall maximum average enrollment size was seen in 1999 with an average of 541 and the minimum was seen in 2011 with an average of 151. Comparing

enrollment sizes between "Completed" studies and "Not Completed" studies (which includes "Ongoing" studies), the trends of average enrollment sizes remain similar. The highest average enrollment is seen in 1997 for "Not Completed" studies with an average of 621 and the lowest is seen in 2017 for "Completed" studies with an average of 120. The size of the bubbles for completed studies are generally smaller compared to the bubbles for the "Not Completed" studies, especially in the later years, showing smaller enrollment sizes in completed studies as the years progress.

2.3.5 Surgical Categories

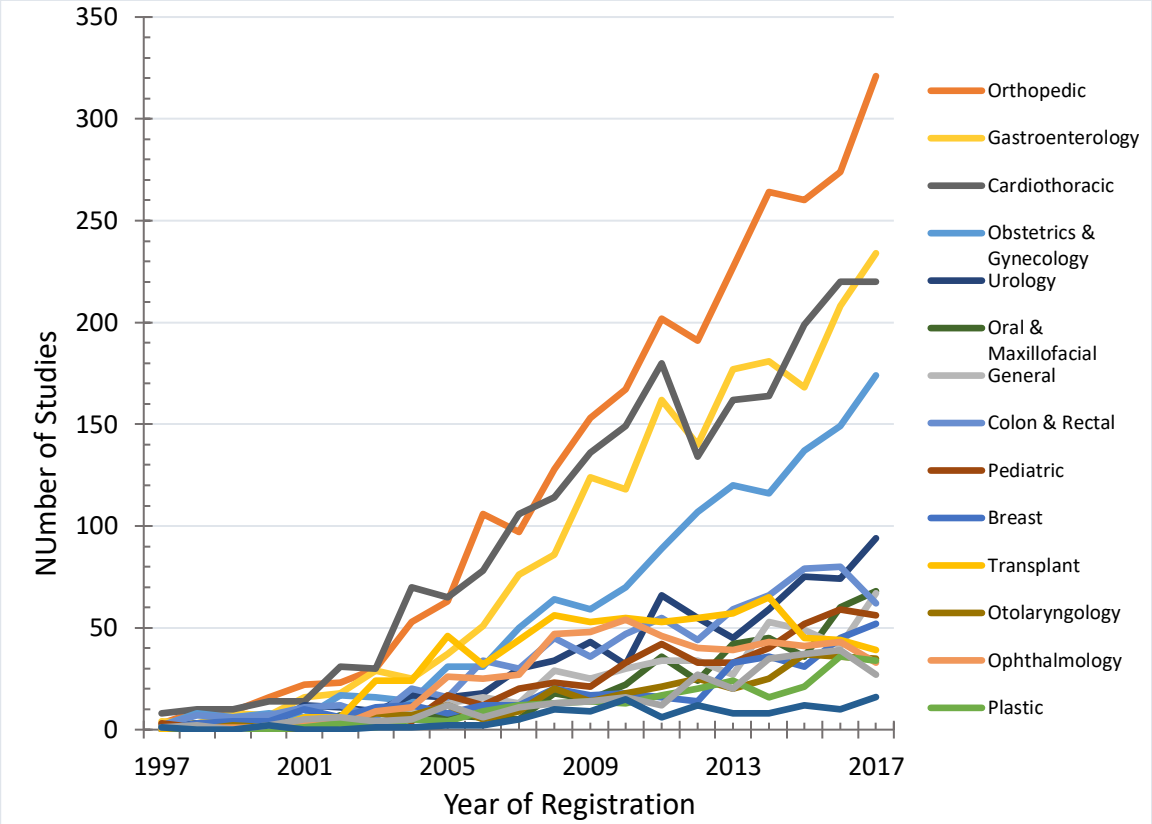


Figure 4: Surgical categories from 1997 to 2017

Regarding the distribution of surgical categories, the four most common categories are Orthopedics (20.0%), Cardiac, Vascular and Thoracic (16.0%), Gastroenterology (14.1%), and Obstetrics and Gynecology (9.6%) (Table 1) (Figure 4). These 4 categories also increase at a faster rate over the years than the others (Figure 4).

It was not until 2003, 6 years into the data, that all the surgical categories were present. Prior to that, at least 1 surgical category was not represented in the data each year. A list of specific surgical procedures and their corresponding surgical categories can be found in Appendix 4.

2.3.6 Geographic Distribution of Registered Surgical RCTs

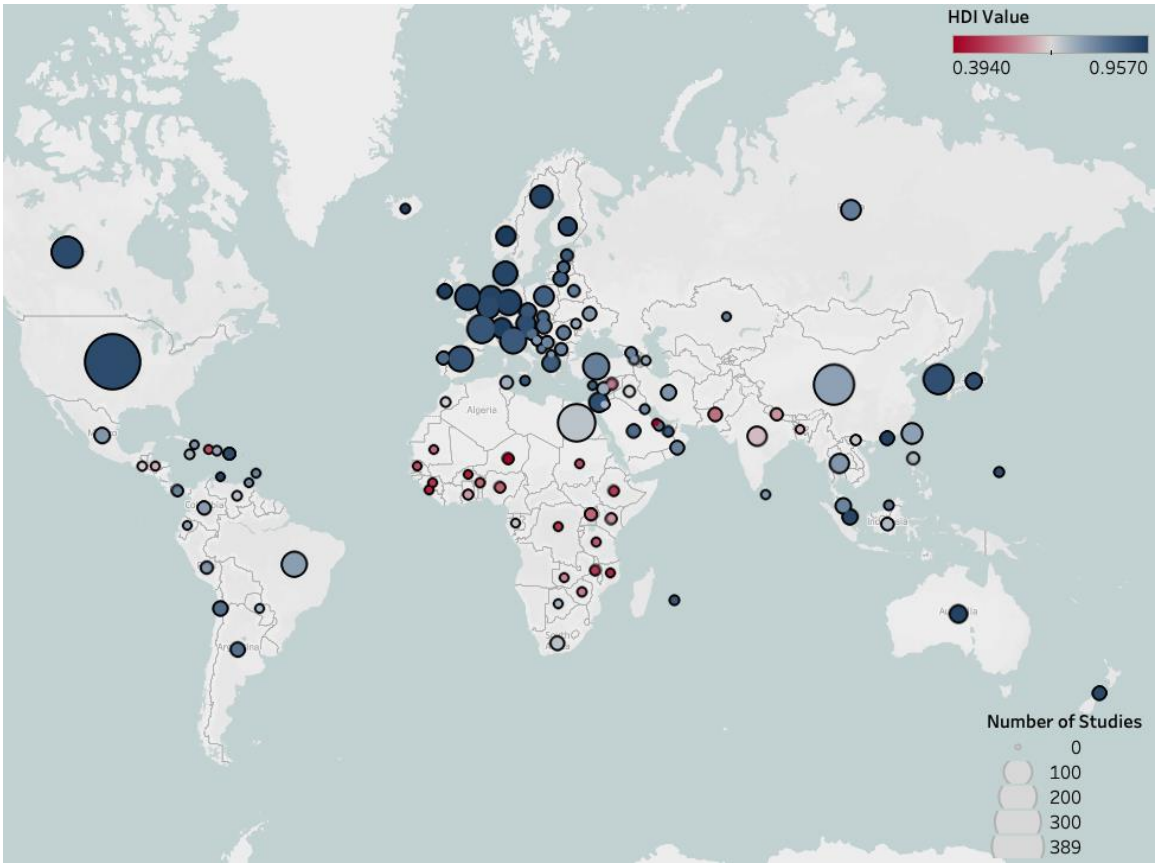


Figure 5: Number of surgical RCTs registered according to country Human Development Index.

The size of the bubbles indicates the number of studies, while the colour of the bubbles indicates the HDI of the country, where the dark red to dark blue colour gradient represents low to very high HDI countries, respectively.

In Figure 5, the geographic distribution and HDI status indicates that most registered studies took place in very high-income countries. Studies from high HDI had larger enrollment sizes compared to registered studies from low-income countries. In

total, 76.8% of registered surgical RCTs were from very high-income countries, mostly in North America, Europe, and some parts of Asia. In contrast, only 0.3% of registered surgical RCTs were from low-income countries.

2.4 Discussion

In total, 13,761 surgical RCTs were registered in ClinicalTrials.gov between 1997 to 2017, with a median enrolment size of 96. While the number of surgical RCTs registered per year has increased over the past 20 years, the average enrolment size has decreased. A total of 43% of studies did not reach completion, with an increasing trend of non-completion over the past 20 years. For studies that reached completion, the median time to completion was 25 months (mean: 34 months). The surgical specialties most commonly addressed in registered RCTs included orthopedics (20%), gastroenterology (14%), cardiac, vascular, and thoracic (16%), and obstetrics and gynecology (9.6%). More than 94% of studies are from high-income countries (very high and high HDI), and very few studies are from lower income settings where the majority of the world population resides.

This analysis provides implications for relevance and integrity of the surgical evidence base and provides areas for future improvement. While it is positive that there is a relatively large number of registered surgical RCTs, a significant portion remain incomplete, and this gap between registered and completed seems to be expanding as the years progress. Proportion of completed studies has been decreasing across the years, and especially after 2013. To some degree, this is expected as there is less time in recent years for registered trials to reach completion status; however, we tried to minimize this impact given the median time to completion was about 2 years (25 months). This raises the question, what is the barrier to hinder these studies from completion? Future studies should explore the barriers and facilitators to study completion.[17]

The proportion of studies belonging to areas with a very high HDI also contributes to concerns regarding generalizability and applicability of the evidence base. Registered studies from medium and low HDI countries first appear in 2001 and 2007,

respectively, (each focused on Obstetrics and Gynecology surgery). The combined proportion of 1.6% of registered RCTs from low- and middle-income countries shows clear lack of proportional representation across the HDI categories. Although this is evidently a generalizability issue, it is congruent with the challenges encountered in these settings where there is reduced surgical volume per capita and limited access to surgical and research infrastructure.

Table 3: Distribution of registered surgical RCTs by HDI category compared to world population

HDI Category	World Population[25]	Registered Surgical RCTs
Very High	19.9%	76.8%
High	37.4%	17.3%
Medium	29.9%	1.3%
Low	12.8%	0.3%

The lack of representation from countries with medium and low HDI, and those located in regions other than Europe and the Americas is evident, with an overwhelming percentage in very high HDI settings (76.8%). According to 2020 data from the World Population Review, only 19.9% of the world's population lives in countries with a very high HDI, whereas 37.4% live in high, 29.9% live in medium, and 12.8% live in low HDI countries.[25] The ratios of where the majority of the world population lies is not proportional to where most of the studies are being conducted (Table 3). The global burden of diseases and conditions warranting surgery is not met and necessities of types of surgeries and proportions for populational burdens is unbalanced.[32] The disproportionality of involved countries and respective HDIs provides urgency for study conduct in regions where most of the world's population matches that disease burden.

Average enrollment sizes were seen to decrease over the years, along with smaller enrollment sizes in "Completed" studies in comparison to "Not Completed" studies. Several explanations could be used to describe this trend, providing explanation as to why the decrease in enrollment sizes is not a major concern, while the discrepancy between sizes for "Completed" and "Not Completed" remains to be an issue of integrity.

Questions arise around why the completed studies are smaller, and whether they are smaller than their intended enrollment sizes on ClinicalTrials.gov, which was beyond the remit of this study to explore. Regarding the overall decrease for one, an increased understanding of sample sizes over time allows us to transition to using smaller samples while still provide a similar confidence of results. This could suggest improved understanding of sample size and adequate power in more recent years, allowing for smaller samples to be used. Further, large surgical trials assessing general questions have likely been undergone earlier in the years, and it would be unethical to reproduce these trials as we have a strong understanding of the results. Accordingly, esoteric demands and subdomains of these general matters are more likely the question at hand for the studies conducted in more recent years which don't affect the general population rather a smaller, more niche subgroup which allows for a smaller sample size.

2.4.1 Limitations

Since we had to rely on the information provided in the ClinicalTrials.gov registry for information on the studies, some information may have been incomplete, especially if the trial registry authors failed to enter the correct information or did not update the registry to reflect changing status over time. It was assumed that these registries were up to date and accurate for analysis purposes, and the result of our study needs to be interpreted in this light. Similarly, due to the volume of studies, it was not feasible to go through all 13,761 and check publication status; therefore, we relied on the status posted within the registries for this information as well. In this case "completion" does not necessarily mean "published", but rather that the studies ended normally, and participants are no longer being examined or treated.

2.4.2 Conclusion

While registered surgical RCTs are increasing over the past 20 years, there remains a substantial proportion that are unpublished. Implications and concerns for research integrity are present,, but without further analysis on the unpublished work it is

difficult to make further assumptions.[17] Additionally, the staggering proportion of studies within very high and high-income countries raises concern for the generalizability and applicability of the evidence urging future research to resemble more balanced proportion of study production and surgical type in relation to where the world's population lies. This is an issue to be addressed by policy makers, research funders, journals, and related institutions to help increase the range in surgical research and enhancing proportional capabilities for producing research.

After the totality of global surgical research has been characterized, it is necessary to further explore the publication status of these studies, and information within these publications. In Chapter 3, we will take a sample of the registered studies and search for publications in order to identify the rate of publication. Additionally, we will compare outcomes from the registries to the publication to analyze outcome distortion, spin, and the overall integrity of the studies.

Chapter 3

3 Categorizing and Describing Spin in Surgical Research

High quality research, including randomized controlled trials (RCTs), is necessary to inform decisions about what works in clinical research.[8] However, incentives to distort research and spin the results into positive outcomes to produce results that seem more interesting threatens the integrity and validity of research.[15] We obtained a sample of surgical RCTs registered in ClinicalTrials.gov between 1997 to 2017 in order to characterize completion rate, publication rate, and frequency of positive results. Additionally, this allowed us to quantify detectable distortion and spin in the presentation of the published results in order to characterize the integrity of the evidence base and identify opportunities for improvement.

3.1 Introduction

In a high stakes field such as surgery, research conducted and reported with integrity and transparency is essential since evidence-based decisions require access to complete and unbiased information. Randomized controlled trials (RCTs) are essential to inform decisions about which surgical interventions work, and well-conducted RCTs are generally considered the highest level of evidence. However, pressure to publish and tendency to report more interesting results has led to a significant tendency to 'spin' the results to increase likelihood of publication and to garner attention from readership and the media. These pressures may incentivize subtle efforts or even outright blatant efforts to manipulate results in order to display a positive, or statistically significant, outcome.[15]

Spin can be defined as the deceptive use of language to either mask unfavourable results, place an emphasis on favourable results, or state misleading conclusions often with causal language suggesting something that is not supported by the results.[16] Similarly associated with misleading behaviours, **outcome distortion** refers to numerical results or analyses where negative results are selectively reported or omitted, as well as the replacement of non-significant outcomes with other significant ones such as in the

case of a primary outcome demotion substituted with a significant one, or promoting significant over non-significant results.[17] The presence of either of these alterations jeopardizes integrity at the individual trial level, and may bias the evidence base as a whole, through overestimation of positive results and under-reporting of negative results.

The consequence of these manipulations includes post-hoc biases known as **publication bias** and **citation bias**. Both biases measure the effect of falsely interpreting and presenting the results in order to be perceived as more positive than they are (beyond what is supported by the data).[17] **Publication bias** occurs when positive studies are over-represented in the published literature, and negative studies are more likely to be missing (failure to publish at all, or longer time to publication), due to a variety of interdependent forces that incentivize authors and journal editors to favour positive findings.[18]

Citation bias is the notion that positive studies are more likely to receive more citations, thus making them more discoverable to the public.[17] This can be analyzed either through the number of citations received by a study or by the impact factor of the journal of publication, as more credible journals tend to have a higher visibility. These factors correspond to the number of citations a journal receives, which is quantified by the **Journal Impact Factor (JIF)** using scientometrics to assess the visibility of a journal. When visibility of positive results is heightened through greater propensity to be accepted in a high-impact factor journal and greater propensity to be cited, this further reinforces positive results to be overestimated in the literature.[17]

A main motivation for manipulation of the results is publication in a credible journal. Publication is rewarding to researchers and their affiliated institution as it demonstrates academic achievement, increases credibility, and in turn, leads to potential increases in funding and future opportunities.[13] To reduce publication bias and outcome distortion, prospective registration of clinical trials has been proposed to ensure a priori declaration of study details to improve accountability to maintain their intentions, or transparently announce changes.[10]

Registration for all clinical trials of drugs, biological products or medical devices (excluding feasibility trials) has been made a legal requirement from September 27, 2007 and onwards to be eligible for publication in select high-impact journals according to the Food and Drug Administration Amendments Act, and similar guidelines are followed by other regulatory bodies.[12] Even with this mandate, annual the rate of publication is increasing at a rate of about 2-3% in the mid 1900s, and increasing to 8-9% rate approaching 2012.[14] The ratio of positive versus negative studies in the literature is also growing rapidly.[26] A 2017 study reported that the proportion of positive results in 1990-1991 was 70.2%, increasing to 85.9% in 2007.[26] While positive findings provide vital information, negative or null findings are equally important, and all results should be published without bias preference toward positive or negative findings. Negative implications of research integrity and concerns for validity are evident and must be addressed.

Overreporting of positive results leads to biased estimates of the true effects of interventions, since the best estimate of the 'true' effect requires all information to be available from positive and negative studies so that a balanced estimate of net effect can be provided. If positive results are preferentially cherry-picked and published, an objective estimate of net effect is not possible. Biased estimates may be further reinforced through meta-analyses that synthesize "all" discoverable studies, since negative studies are less likely to be published, and if they are, they may be published after a longer time delay than positive studies. Negative studies are equally necessary for researchers, doctors, and policymakers to support objective evidence-based decision making.[26]

A number of efforts have been advocated to mitigate these biases, including mandates for trial registration to ensure accountability, pre-registered reports to incentivize publication of results without favouring positive over negative research outcomes, and meta-research initiatives to detect biases in the reporting of research.[17] Despite the increased attention to these risks to research integrity, greater efforts are required to better characterize, quantify, and report the net impact of these biases. The aim of this study is to identify and characterize overall bias including failure to publish,

outcome distortion, and spin within a sample of surgical RCTs over the past two decades, and to determine whether there were significant differences between studies with positive versus negative conclusions.

3.2 Methodology

From the full cohort of surgical RCTs registered in ClinicalTrials.gov between 1997-2017 (n=13,761) previously described in Chapter 2 (refer to Chapter 2, Methodology, Section 2.2.), we obtained a sample (n=5,094) and attempted to identify all publications matching the registered studies in order to quantify publication rates, from which we calculated failure to publish. Since it was not feasible to extract data from all published studies, we used sampling techniques from this cohort of studies to identify a representative subset of studies with matching publications to enable in-depth analysis of extracted data to quantify selective outcome reporting and spin in the published studies. Detailed methods are provided below.

3.2.1 Searching for Published Articles to Quantify Non-Publication Rate

To inform the publication rate for the sample of studies (n=5,094), we first limited the registered surgical RCTs to the set that had begun enrolment (n=4,652), and for these, we searched for the corresponding published article(s) within the registered RCTs recorded on ClinicalTrials.gov, as many investigators will list the relevant publications there upon completion. However, since many investigators fail to list the publications stemming from the registered research, it was also necessary to search the medical literature more extensively. To match the studies that did not have their corresponding publications included in the registry, we used a custom Python program (Appendix 5) to find the best-matched articles on PubMed and PubMed Central based on the available information provided in the registration on ClinicalTrials.gov. If the program was unable to identify the publication(s), all relevant criteria from the registry (including NCT identifier, official title, authors, other study identifier, intervention) were used to manually search Medline and Google Search for any relevant articles. If publication(s)

could not be found through using the custom Python program and a minimum of 5 minutes of manual searching had elapsed, it was deemed that the study was not published at the time of searching. Matching the studies to their corresponding publications took place between mid-November 2020 to the end of July 2021. Using the publications identified, we first quantified the overall **publication rate** for the sample of registered RCTs, and also used this to quantify **failure to publish**.

3.2.2 Sampling Methods to Identify a Subset of Published Studies for In-Depth Review

Due to the large corpus and time constraints, a sample of studies was selected for review. Within this sample, the identified published studies were further randomly sampled for data extraction. However, to ensure that we had an even distribution of studies from each year, we first performed an intentional selection with 10 studies consisting of the 5 highest and 5 lowest enrollment sizes as reported by the research on their registry. After the intentional selection was complete, the following rounds each consisted of randomly selected studies proportional to the number of studies registered in that year, using 2.75% of the registered studies per year each round. The samples were randomly determined using Microsoft® Excel (v.16.51). This process was repeated until data saturation was reached; that is, the point where no new information is obtained by collecting further data as defined by no meaningful difference in the proportion of studies with selective outcome reporting and spin after data for at least three successive samples of studies had been extracted. To validate our claim of saturation and generalizability, we also compared the baseline characteristics of the subset of studies with that of the original large cohort of studies to check whether the distribution of types of surgical procedures and distribution across human development indices (HDI) and geographic region was similar.

After using sampling techniques to identify a representative subset of registered studies with matching publications, we used the publications for in-depth data extraction to quantify the incidence of outcome distortion and spin (defined below) for all the

studies included in the subset, subdivided by studies with positive and negative results, separately.

3.2.3 Data Extraction

Data extraction for characteristics available in the registry were described in Chapter 2, Section 3.2, Methodology. In addition, for this chapter, we examined other characteristics of the clinical trials, including enrollment size, outcomes, indicators to inform spin, indicators to inform outcome distortion, and their impacts including number of positive studies, citations, journal impact factors, and time to publication. See Appendix 7 for the extraction guidelines for the categories and their definitions. Data extraction occurred between January 2021 to July 2021.

3.2.3.1 Outcome Definitions – Publication Rate, Failure to Publish, Outcome Distortion, Spin

Key outcomes of interest included proportion of registered studies that were published, and of those, how many were positive, and turned positive due to outcome distortion or spin. Statistical comparison of characteristics associated with positive studies versus negative studies was planned in order to assess whether positive studies were associated with outcome distortion, spin, and shorter time to publication.

Publication rate was defined as the total number of RCTs registered on ClinicalTrials.gov between 1997-2017 with at least one identifiable publication, as of July 2021. **Failure to publish** was defined as (1 – publication rate).

Outcome distortion was defined as the presence of one or more of the following questionable reporting practices relating to outcomes when comparing the original study registration versus the publication, in alignment with the definition previously provided by Devries et al: [17] removal of non-statistically significant primary outcomes, addition of statistically significant primary outcomes, changes in planned time frame for the primary outcome, or omission of intended primary outcomes entirely. To measure distortion, we

compared up to the first five primary outcomes reported in the RCT registration on ClinicalTrials.gov (which we refer to as intended primary outcomes) to the outcomes actually reported in their publication (which we refer to as reported primary outcomes). Specifically, we recorded the direction (positive or negative), effect measure, and significance of the results (using p-values, confidence intervals, or written suggestion if other indicators were not identified). We also recorded whether the intended primary outcomes were adequately defined and if the reported primary outcomes were clinically relevant. Secondary outcomes were also obtained from the registry and the published studies. However, detailed information on significance and effect measures were not recorded for secondary outcomes, and due to time limitations, we focused only on alterations of primary outcomes. Further details can be found in Appendix 6 and Table 4.

Spin is a "specific reporting strategy" involving result misrepresentation that can "highlight that the experimental treatment is beneficial, despite a statistically nonsignificant difference for the primary outcome, or to distract the reader from statistically nonsignificant results".[15] In this study, we focused on spin in the conclusion, where the overall interpretation and presentation of the outcomes was addressed. To measure spin in the studies, quotes of the concluding statement from the abstract and the discussion section were retrieved and categorized into positive or negative outcomes, where a negative outcome suggests that there was no significant difference in the interventions, and a positive outcome suggests that there was a difference. This allowed us to determine whether the study results were turned positive if the primary outcomes were actually non-significant. Along with this, levels of spin in the conclusion were classified as low, medium, or high as per definitions from Boutron and colleague's classifications and definitions of spin (Appendix 6).[15] Refer to Table 4 below for details.

Table 4: Definitions for outcomes distortion and spin

Category	**Outcome Distortion**	**Spin**
Definition	The presence of questionable outcome reporting practices when	Misrepresentation of results to "highlight that the treatment is beneficial despite a non-significant

	comparing the registration versus the published study. Examples include **switching primary outcome to achieve significance, or downplaying preplanned primary outcomes when non-significant, or revising the definition of the primary outcome or its timepoint to achieve significance.**	difference for the primary outcome, or to distract the reader from statistically non-significant results."[15] Note that this differs from Outcome Distortion since it is an assessment of the appropriateness of the authors' ***interpretation*** **of the results as communicated in the published conclusion**.
What was compared	Changes in primary outcomes intended (from the RCT registration) versus primary outcomes actually reported (from RCT publication)	Overall conclusions declared in the published article versus the underlying supporting evidence.
Details for what was extracted	Primary outcomes intended and their significance, primary outcomes reported and their significance, and information to determine whether any of the following occurred: - a non-significant primary outcome was demoted - a significant secondary or other outcome was promoted to primary -intended primary outcomes were entirely missing - time frames for outcomes were different from intended	Quotes from abstract conclusion or main manuscript conclusion, followed by judgement of whether this was considered low, moderate, high spin or no spin at all according to Boutron definitions: [15] **No spin: "**the conclusion is consistent with the results and highlights adverse events" **Low spin:** "acknowledgment of the statistically nonsignificant results for the primary outcome OR uncertainty and recommendations for further trials"

		Moderate spin: "no acknowledgment of the statistically nonsignificant results for the primary outcome AND uncertainty or recommendations for further trials" **High spin:** "no acknowledgement of the statistically nonsignificant results for the primary outcome AND no uncertainty AND no recommendations for further trials" (Appendix 6).[15]
Where it was found	Results section	Abstract conclusion or main manuscript conclusion

3.2.3.2 Definition for Impacts of Distortion: Positive Studies, Citations, Journal Impact Factor, and Time to Publication

To further characterize measurable impacts of **publication bias, outcome distortion, and spin**, we analyzed the **proportion of positive studies**, the **number of citations** a study received, the **journal impact factor**, and **time to publication**.

A **positive study** was defined as one where the conclusion declared a significant difference between interventions. A **negative study** was defined as one where the conclusion did not declare a significant difference between interventions. We determined whether a study concluded positive or negative results from the abstract conclusion where the overall difference between interventions was reported. In the rare case that there was no abstract, the discussion conclusion was used to classify the study as positive or negative based on the description of the overall study results.

The **proportion of positive studies** was calculated by taking the number of positive studies and dividing it by the total number of published studies. The **proportion of studies that turned positive** was calculated by taking the number of studies that turned positive and dividing it by the denominator which is the number of all studies. Studies could "turn positive" either through distortion of outcomes, or through spin in the conclusion. As a sensitivity analysis, the **proportion of positive studies that were turned positive** was also calculated by taking the number of studies that turned positive as the numerator and dividing it by the number of positive studies, which we set as the denominator.

The **number of citations** for each published study was retrieved from Google Scholar. The **journal impact factor** was defined as that provided by Clarivate Journal Citation Reports page according to the year of study publication.[27] If the journal impact factor could not be obtained, then it was defined as unknown. Additionally, since at the time of data extraction, the journal impact factor for the year 2021 had not been released. Therefore, for studies published in 2021, the 2020 JIF was assigned since there is limited variability from year to year.

Time to publication was determined by taking the duration (in months) between the date of study registration on ClinicalTrials.gov and the date of online publication. If the full date was not available, the first day of the month was taken or the first day of the year if the month of publication also was not found.

In the case that multiple publications resulted from a single trial registration, the first publication was preferentially extracted for descriptive characteristics. However, all relevant publications were examined for comparing between the outcomes declared in the registration versus those actually published, and all relevant publications were explored for the significance analysis. This was to ensure that longer time frames were considered for outcome reporting bias purposes as authors often published multiple time frames prior to their longest intended one.

3.2.4 Statistical Analysis

Microsoft® Excel (v.16.51) was used for descriptive statistics. Statistical inferences were calculated using RStudio (Version 1.2.5019).[28] Chi-square test was performed for categorical outcomes, while Welch's two sample t-test was performed for continuous data, where a two-sided $p<0.05$ was considered statistically significant.

An icon array was created on Infogram.com to show the proportion of studies that progressed from registration to publication. The stages of this progression included all surgical RCTs registered between 1997-2017, studies that began enrollment, and studies that were published. Studies that were registered were analyzed from the sample of studies that were evaluated to see if the published version could be found. Studies that began enrollment were retrieved from ClinicalTrials.gov, and if their reported enrollment size was larger than 0, it was assumed that enrollment had begun. We created a second icon array to show the degree of outcome distortion and conclusion spin compared to the intended outcome from the published studies. A study was considered to have turned positive if the intended primary outcome was either negative, unknown or mixed based on the primary outcomes intended from the registry, and the overall conclusion of the study had reported a positive or a significant conclusion.

Regression analyses were performed using RStudio (v.1.2.5019) to explore the change in positive studies over time.[28] Logistic regressions were performed using a quasi-binomial distribution with year as predictor variable for three categories including the proportion of positive studies, the proportion of studies that turned positive, and the proportion of positive studies that were turned positive. We also explored the proportion of studies that were turned positive by using three different characteristics: studies turned positive due to outcome distortion, due to spin, and overall.

3.3 Results

3.3.1 Included Studies & Publication Rate

In total, 13,761 registered studies were included in the original full database of registered surgical RCTs. Baseline characteristics of these studies are summarized in Chapter 2, Table 1. After 5 rounds of sampling from the original database of 13,761 registered RCTs, saturation was reached, and the resulting subset of 5,094 studies was used for in-depth data extraction. Of the 5,094 registered studies subjected to medical database searches in an attempt to match them to their published results, only 1,718 had published results (Figure 1). This represents a **publication rate** of 33.7% (1,718/5,094), and a **failure to publish rate** of 66.3%.

3.3.2 Study Characteristics

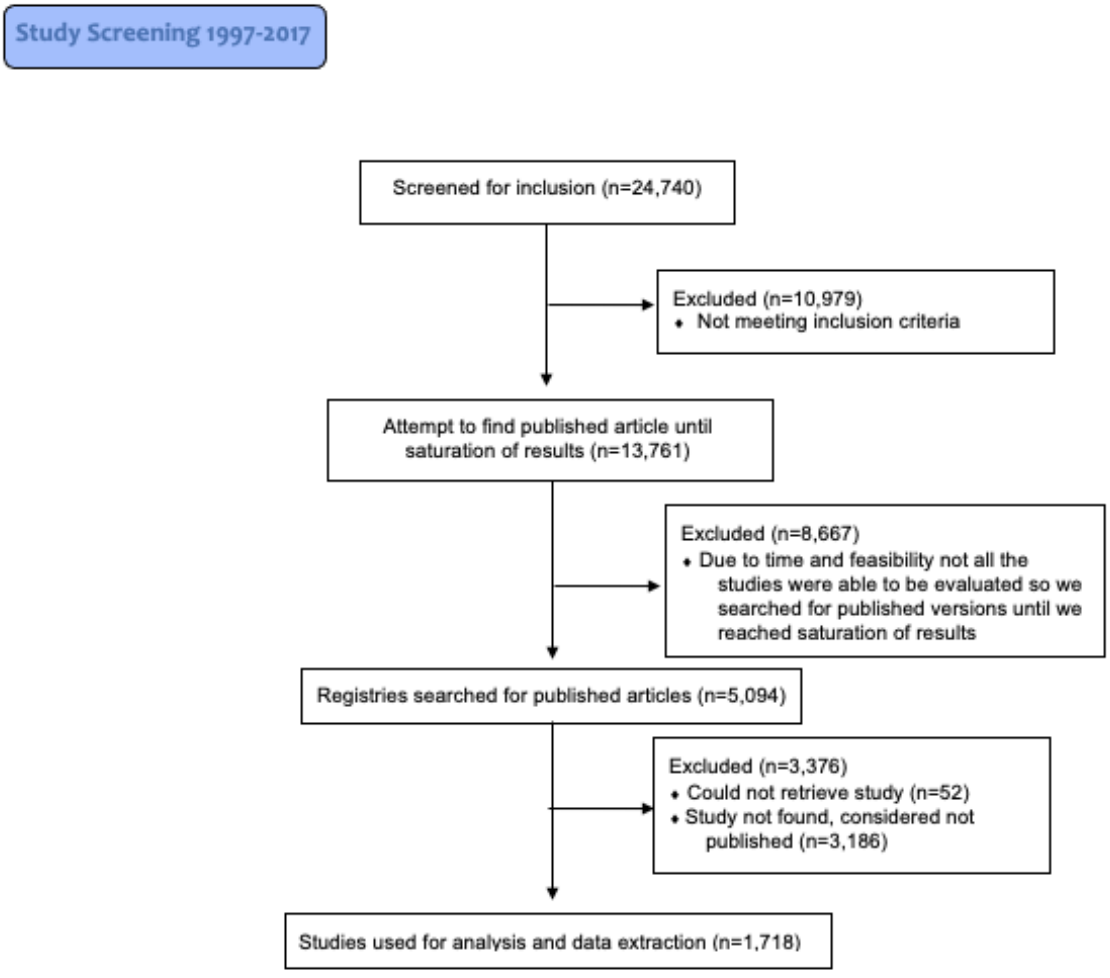


Figure 6: Flow chart of study process from screening to data extraction

Characteristics of this sample of studies in which their publication was found (n=1,718) are summarized in Table 5 and are comparable to baseline characteristics of the original larger cohort of surgical RCTs described in Chapter 2, Table 1. Median enrollment size was 92 (IQR: 51-186) and mean was 322 (range: 5-18,876). The majority of trials were conducted in countries with a very high Human Development Index (HDI) (78.6%) or high HDI (18%) and were largely conducted in Europe (50.2%) and the Americas (27.7%). The four most common surgical categories included: Orthopedics (20.5%), Cardiac, Vascular and Thoracic (14.4%), Gastroenterology (14.2%) and Obstetrics and Gynecology (10.4%).

Since our sample consists only of studies that had been published, it is expected that the status of these studies would be "Completed" on ClinicalTrials.gov. However, only 82% were labelled as such, which indicates failure of some investigators to update their ClinicalTrials.gov entries upon trial completion. Further, 15% of these studies were labelled "Not Completed", and 4% were labelled "Ongoing". Additionally, we found that 14% of studies had inadequately defined primary outcome(s) in their trial registration, which suggests lack of transparency with their intentions during registration of their clinical trial.

3.3.3 Study Characteristics for Positive versus Negative Studies

Out of the 1,718 studies retrieved for data extraction, 1,058 (62%) were positive studies and 660 (38%) were negative studies. The proportion of positive studies differed significantly across surgical categories (p=0.018) with the highest being Orthopedic for both positive (19.5%) and negative (22.1%) studies, and the lowest being Thyroid for both positive (1.5%) and negative (0.5%) studies. Mean enrollment size was significantly larger for negative studies (p=0.069).

The Western Pacific Region had the highest proportion of positive studies at 70%, whereas the African Region had the lowest proportion at 40%. Statistical inferences were

not made for geographic regions since many studies had multiple regions and it was difficult to attach a region to a single study. The proportion of positive studies was similar across HDI categories (p=0.62), and across completion status ("Completed", "Not Completed", and "Ongoing") (p=0.15).

Table 5: Characteristics of published studies

Characteristic	Data Extracted Sample (n=1,718)			
	Total (n=1,718)	Positive Studies (n=1,058)	Negative Studies (n=660)	P-value
	No. (%)	No. (%)	No. (%)	
Location by Region				
Americas	749 (27.7)	403 (26.6)	346 (29.1)	
Europe	1,356 (50.2)	748 (49.4)	608 (51.1)	
Western Pacific	341 (12.6)	202 (13.4)	139 (11.7)	
Eastern Mediterranean	160 (5.9)	112 (7.4)	48 (4.0)	
South-East Asia	67 (2.5)	36 (2.4)	31 (2.6)	
Africa	30 (1.1)	12 (0.8)	18 (1.5)	
HDI[1] Category				
Very High	1,346 (78.6)	797 (75.8)	549 (84.2)	
High	311 (18.2)	224 (21.3)	87 (13.3)	0.62
Medium	40 (2.9)	27 (2.6)	13 (2.0)	
Low	6 (0.4)	3 (0.3)	3 (0.5)	
Surgical Category				
Breast	39 (2.3)	28 (2.7)	11 (1.7)	
Orthopedic	352 (20.5)	206 (19.5)	146 (22.1)	0.018
Neurology	27 (1.6)	19 (1.8)	8 (1.2)	

Transplant	70 (4.1)	42 (4.0)	28 (4.2)	
Obstetrics and Gynecology	178 (10.4)	114 (10.8)	64 (9.7)	
Plastic	33 (1.9)	19 (1.8)	14 (2.1)	
Urology	82 (4.8)	48 (4.5)	35 (5.3)	
Pediatric	82 (4.8)	58 (5.8)	24 (3.6)	
Cardiac, Vascular and Thoracic	247 (14.4)	138 (13.0)	109 (16.5)	
Otolaryngology	36 (2.1)	18 (1.7)	18 (2.7)	
Thyroid	19 (1.1)	16 (1.5)	3 (0.5)	
Dental, Oral and Maxillofacial	88 (5.1)	59 (5.6)	29 (4.4)	
Colon and Rectal Surgery	103 (6.0)	69 (6.5)	34 (5.2)	
Ophthalmic Surgery	67 (3.9)	38 (3.6)	29 (4.4)	
Gastroenterology	244 (14.2)	163 (15.4)	81 (12.3)	
General	51 (3.0)	23 (2.2)	27 (4.1)	
Completion Status[3]				
Completed	1,403 (81.7)	865 (81.8)	538 (81.5)	0.15
Not Completed	250 (14.6)	160 (15.1)	90 (13.6)	
Ongoing	65 (3.8)	33 (3.1)	32 (4.9)	
Enrollment Size, median (IQR)	91.5 (51.0-186)	86 (50-155)	101 (56-242)	
Enrollment Size, mean (range)	321.6 (5-18,876)	235.8 (5-13,698)	459.8 (9-18,876)	<0.0001

1. HDI = Human Development Index
2. Not applicable

3. Author reported from ClinicalTrials.gov registry

3.3.3.1 Outcome Distortion

Incidence of outcome distortion in studies was 81% overall and was not significantly different for positive versus negative studies (82% vs 80%; p = 0.37). Incidence for the either outcome distortion, inadequate defining of outcome, or clinically irrelevant outcomes was 91% total, with no significant difference for positive versus negative studies (92% vs 87%, p=0.12). Sub-categories of type of outcome distortion are outlined in Table 6. Overall, 14% of studies inadequately defined their primary outcomes, 11% had non-clinically relevant primary outcomes, and 80% had a discrepancy between outcomes intended and actually reported. In addition, 10% demoted a non-statistically significant primary outcome to secondary outcome, 14% promoted a significant secondary outcome to primary outcome status, 13% did not report the primary outcome at all, and 11% reported their primary outcome with a revised time frame. In total, 14% of studies "turned positive" due to switching the originally planned primary outcome to another significant outcome.

The sub-categories of outcome distortion that were significantly higher for positive versus negative studies included non-clinically relevant primary outcomes (12% vs. 8%, p=0.017), a statistically significant outcome promoted to primary status (20% vs. 3%, p<0.00001), and "turned positive" due to switching the originally planned primary outcome to another significant outcome (20% vs. 3%, p<0.00001).

3.3.3.2 Spin

Incidence of spin was 39% overall and was significantly greater for positive versus negative studies (45% vs 30%, p <0.00001). 'High spin' was found in 7% of all studies and was significantly greater for positive versus negative studies (12% vs 0%, p <0.00001). 'Medium spin' was found in 11% of studies overall and was significantly greater for positive versus negative studies (18% vs 0.2%, p <0.0001). Low spin was found in 21% of studies and was significantly greater for negative studies versus positive

studies (30% vs. 15%, p<0.00001). Similarly, no spin was also significantly greater for negative studies (70% vs. 55%, p<0.00001). Results turned positive due to spin in the conclusion or selective outcome reporting in the conclusion in 16% of studies and was significantly greater for positive versus negative studies (27% vs 0%, p <0.00001).

3.3.3.3 Citation Bias & Journal Impact Factor

Mean number of citations did not differ significantly between positive and negative studies (68 vs 81; p=0.26). Mean JIF was significantly lower for positive compared with negative studies (p<0.0001). In addition, positive studies had significantly more missing JIFs (p=0.0068). Positive studies also had shorter time to publication compared to negative studies (53 vs 67 months; p<0.0001).

Table 6: Statistical comparison of positive versus negative studies

Characteristic	Total (n=1,718)	Positive Studies (n=1,058)	Negative Studies (n=660)	P-value
	No. (%)	No. (%)	No. (%)	
Inadequately defined primary outcomes	242 (14.1)	163 (15.4)	79 (12.0)	0.071
Non-clinically relevant primary outcomes (reported)	184 (10.7)	131 (12.4)	53 (8.0)	0.017
Outcome distortion (at least 1 incidence)	1,398 (81.4)	868 (82.0)	530 (80.3)	0.37
Discrepancy in outcomes reported and outcome intended (primary and secondary)	1,373 (79.9)	848 (80.5)	525 (79.5)	0.76
Non-statistically significant primary outcome demoted	169 (9.8)	99 (9.4)	70 (10.6)	0.40

Statistically significant outcome promoted to primary	233 (13.6)	215 (20.3)	18 (2.7)	<0.00001
Intended primary outcome not reported at all	225 (13.1)	141 (13.3)	84 (12.7)	0.72
Primary outcomes were the same but with different time frames	190 (11.1)	122 (11.5)	68 (10.3)	0.43
Study “turned positive” due to primary outcomes yielding statistically significant results when intended had either non-statistically significant or unknown	232 (13.5)	213 (20.1)	19 (2.9)	<0.00001
Spin (any)	674 (39.2)	476 (45.0%)	198 (30.0%)	<0.00001
High Spin	124 (7.2)	124 (11.7)	0 (0)	<0.00001
Medium Spin	193 (11.2)	192 (18.2)	1 (0.2)	<0.00001
Low Spin	356 (20.7)	159 (15.0)	197 (29.9)	<0.00001
No Spin	1,044 (60.8)	582 (55.0)	462 (70.0)	<0.00001
Results “turned positive” due to spin in the conclusion or selective outcome reporting in the conclusion	280 (16.3)	280 (26.5)	0 (0)	<0.00001
Citation bias/Publication bias				

Number of Citations, median (IQR)	16 (4-47.8)	16 (4-47)	15 (5-49)	
Number of Citations, mean (range)	73.3 (0-3,024)	68.3 (0-2413)	81.2 (0-3024)	0.26
Journal Impact Factor, median (IQR)	3.5 (2.3-6.0)	3.4 (2.2-5.4)	3.7 (2.4-6.9)	
Journal Impact Factor, mean (range)	7.9 (0.09-91.3)	6.4 (0.09-91.3)	10.2 (0.23-91.3)	<0.00001
Journal Impact Factor, number not available	237 (13.8)	159 (9.3)	78 (4.5)	0.0051
Time to Publication (months) median (IQR)	50 (35-73)	46 (33-65)	58 (40-86.3)	
Time to Publication (months), mean (range)	58.7 (1-226)	53.4 (1-224)	67.1 (8-226)	<0.001

3.3.4 Regression Results

Table 7: Logistic regression results for proportion of positive studies, proportion of studies turning positive out of *all* studies, and proportion of studies turning positive out of the *positive* studies

Proportion	Type of Distortion	Coefficient		
		Beta Estimate	Standard Error	p-value
Positive Studies	NA[1]	0.040	0.0131	0.0066
Studies that turned positive out of all studies	Due to outcome distortion	0.0033	0.016	0.84
	Due to Spin	-0.028	0.014	0.056
	Overall	-0.0092	0.013	0.51
Positive studies that turned positive out of positive studies	Due to outcome distortion	-0.021	0.020	0.30
	Due to spin	-0.062	0.018	0.0028
	Overall	-0.046	0.018	0.018

1. Not applicable for the proportion of positive studies

After exploratory logistic regressions shown in Table 7, multiple trends and relationships regarding positive outcomes and studies that have turned positive were observed. Figure 7 shows the increasing proportion of positive studies published each year (Table 4: β=0.040, SE=0.013, p=0.0066).

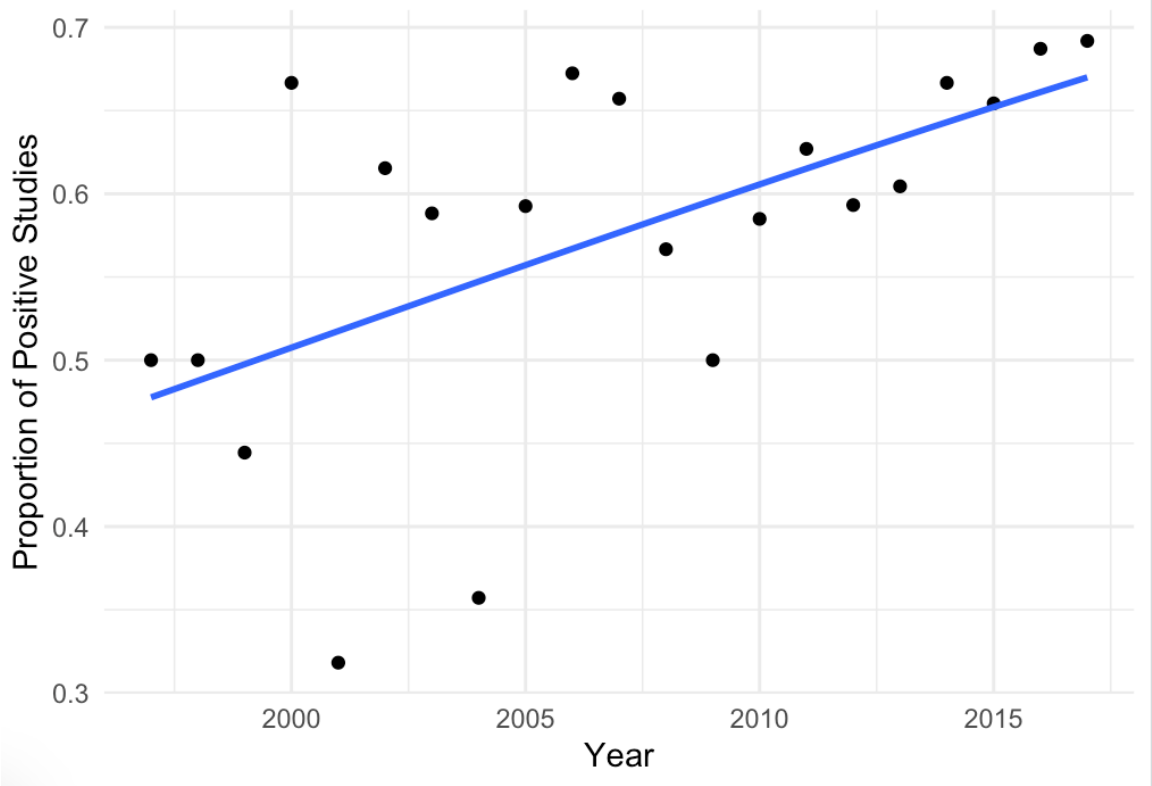


Figure 7: Logistic regression of the proportion of positive studies from 1997 to 2017 (numerator is positive studies, denominator is all studies)

In contrast, Figure 8 indicates that the proportion of studies that turned positive decreased over time (Outcome Reporting p=0.84, Spin p=0.056, Overall p=0.51), and most of this decline could be explained by decreasing incidence of studies that turned positive due to spin. However, none of these was significant.

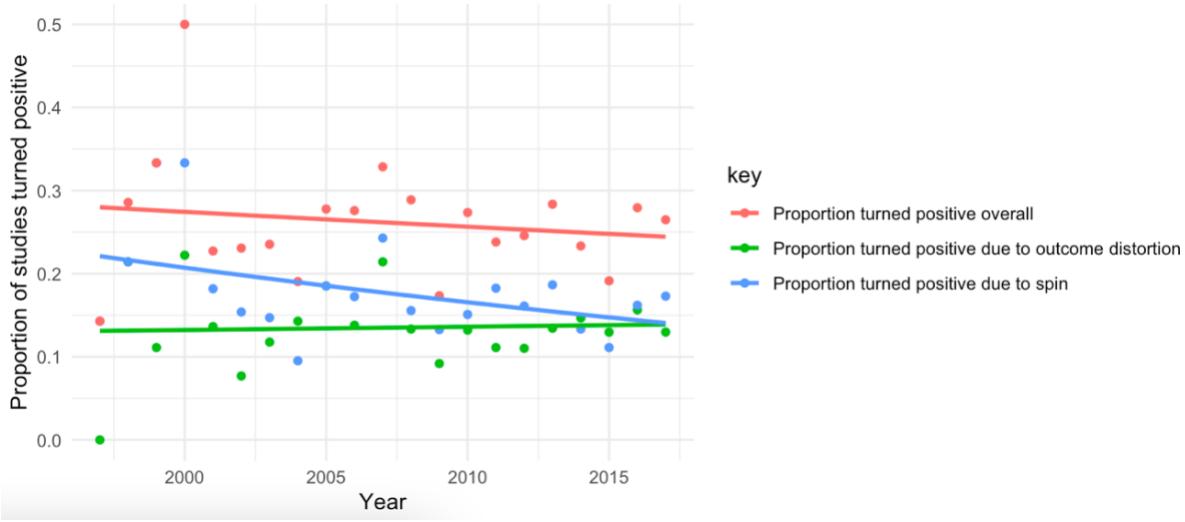


Figure 8: Logistic Regression of the proportion of studies that turned positive from 1997 to 2017 (numerator is studies that turned positive, denominator is all studies)

As a sensitivity analysis, we also explored the proportion of positive studies that were turned positive, where the denominator is the number of positive studies (rather than a denominator of all studies) (Figure 9), and found a significant decline in proportion of positive studies that were turned positive over time (Table 7: spin: β=-0.062, SE=0.018, p=0.0028, overall: β=-0.046, SE=0.018, p=0.018), with most of the decline related to proportion turned positive due to spin.

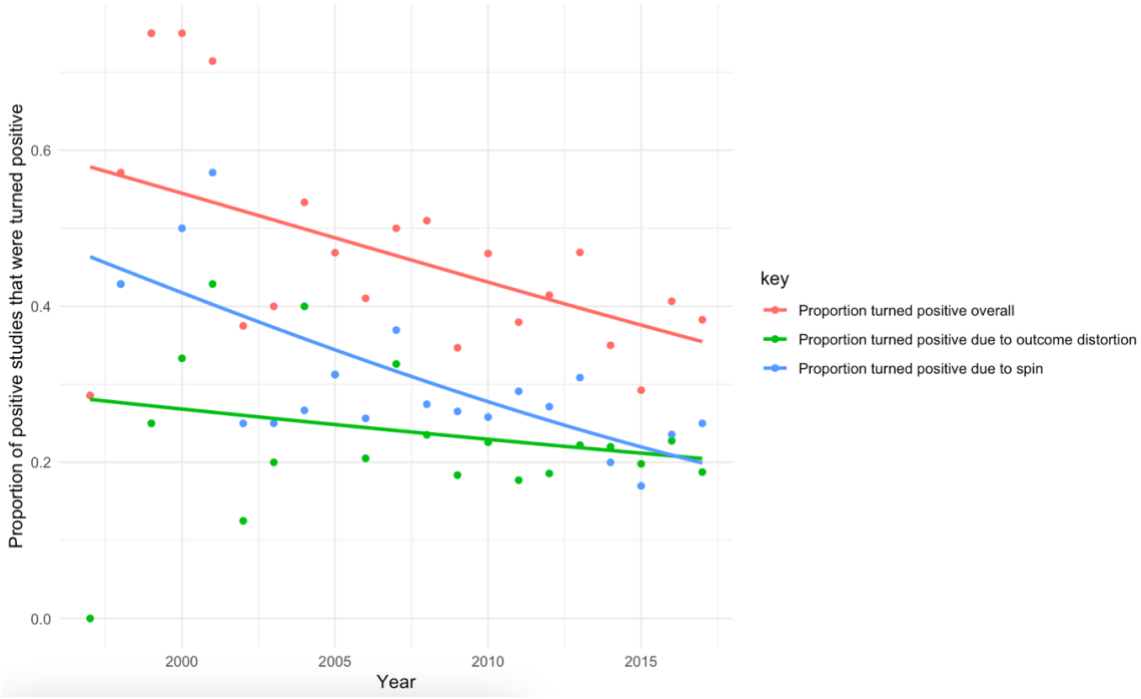


Figure 9: Logistic regression of the proportion of positive studies that turned positive from 1997 to 2017 (numerator is studies that turned positive, denominator is all positive studies)

Together, these regressions suggest that even though there is an increasing proportion of positive studies over the years, the number of positive studies that were *turned* positive due to detectable spin or outcome distortion is decreasing over time.

3.3.5 Icon Array Displaying Sample of Studies Selected to Search for Publication, and Distortion for Published Studies

To summarize the key results of this study into one figure, an icon array was created to visualize the progression of studies starting from the sample of registered surgical RCTs searched for evidence of publication (n=5,094), followed by studies that began enrollment, and studies that were published (Figure 10). Further, from the studies that had been published, the proportion of intended outcomes were displayed followed by the actual outcomes reported, according to whether they were positive or negative studies, and finally, according to the proportion of studies with detectable spin.

Out of the sample of registered studies, only 33.7% were published, which translates to a failure to publish rate of 66.3%. Of the originally intended primary outcomes, 36.2% were positive (622/1,718). However, when analyzing the reported outcomes, there were many more positive studies (61.6%), including a large portion that turned positive from either negative, unknown, or mixed results (25.4%). The studies that turned positive had the most spin detected within their conclusions. In total, the proportion of registered studies that were published without outcomes turning positive or detectable spin was only 18% (918/5,094). If all aspects of outcome distortion were considered, the proportion of studies without outcome distortion or spin would be less than 18% (not calculated in this study, due to non-mutually exclusive definitions for distortion and spin).

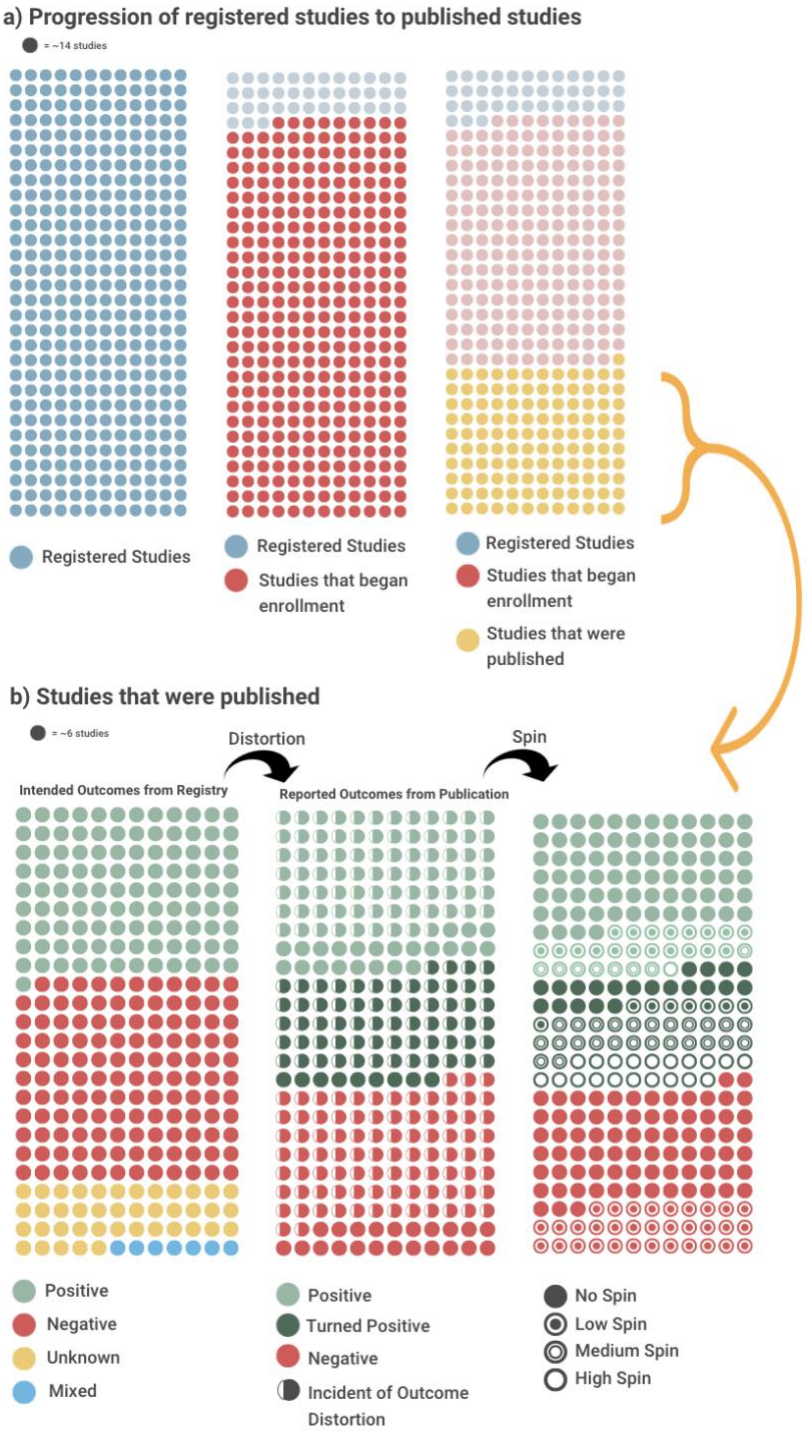


Figure 10: Icon array displaying the proportion of registered studies that were published, and of those, what proportion were positive studies, what proportion had outcome distortion, what proportion turned positive and what proportion had spin

3.4 Discussion

While the number of surgical RCTs registered on ClinicalTrials.gov has increased from 31 to 1,619 per year over the past 2 decades, only 34% of registered studies with patient enrollment made it to publication, which translates to a non-publication rate of 66%. Of the 34% of registered studies that do make it to publication, 81% showed evidence of outcome distortion, and 39% showed evidence of spin in the conclusion (distortion and spin are non-mutually exclusive categories). Altogether, this study provides tangible evidence of the troubling state of the surgical evidence base, a significant proportion is entirely missing or reported with significant bias in order to "try" to achieve significant results through manipulation and questionable reporting research and reporting practices.

This low publication rate of registered RCTs provides direct evidence of publication bias in the field of surgery and is even worse than the publication rate of 46-50% detected in previous studies from other fields in medicine.[29, 30] This raises questions of why more registered studies that are started are not making it to publication. Whether non-publication is due to study attrition (including slow enrolment, investigator fatigue, changes in research priority due to evolving evidence elsewhere), failure to achieve results that fit with preferred conclusions (preferring to suppress negative or unexpected results), or lack of sufficient resources for investigators to complete the study, analyze the results, and follow through to successful publication (publication requires iterative submission cycles and responses to peer reviewer feedback) remains unknown. Further research is required to ascertain the "failure rate" at each stage of research, from registration to publication, in order to better inform how to mitigate research attrition, misrepresentation of results, and overall research waste.

3.4.1 Study Characteristics

While the number of registered studies increased over years, study enrollment size decreased in more recent years. The median enrollment size (92 patients) was smaller

than the mean (322 patients), suggesting that earlier years had a larger enrollment size. Potential reasons for increased studies with smaller size include possibility that a number of large studies of the most prevalent conditions have already been done in earlier years, leading to increasingly esoteric demands for sub-questions and subspecialty conditions affecting a smaller subgroup of individuals. Other explanations might include an increased number of clinical investigators vying for research, preferring to conduct research independently rather than collaboratively.

A wide range of surgical categories was found in our cohort of registered RCTs. However, the Americas and Europe dominated most of the RCT registrations, and almost 97% were from high HDI or very high HDI settings, again raising questions about applicability and generalizability to regions of the world where the majority of the population lives and where most of the unmet burden of disease amenable to surgical intervention exists.

The mismatch between the proportion of studies labelled as ‘completed’ on their ClinicalTrials.gov registry entry from the published sample is a concerning indicator of the reliability of the registries, given that published studies generally indicate ‘completed’ studies. Even though mandates have been put in place to require completed registries in order to publish in certain journals, updating registries is also important so that clinicians and the public can stay informed on changes and progress made in the trials and researchers can avoid duplicating research unnecessarily.

This concern is also amplified by the finding that 14% of RCT registry entries had inadequately defined primary outcomes, suggesting that the study registration is no guarantee of the quality necessary to maintain integrity and transparency in the literature. Lack of adequate definition of outcomes, and failure to keep the registries up to date, represents significant threats to the ability for registries to contribute to their aspired quality improvement without adequate auditing and oversight for accountability between registered plan versus reported study. Without full transparency, there is a temptation to manipulate the data and frame it into whatever is giving the best story. The nebulousness of the data when it is prospectively explained in this format allows for flexible reporting

that is technically still abiding by typical journal publication requirements yet produces ambiguous or misleading results which are amplified across the literature and decision-making.

3.4.2 Comparison of Positive versus Negative Studies

Out of the 1,058 studies reported as positive, only 622 (58.8%) were originally positive before questionable research and reporting practices were applied to achieve a positive outcome. In particular, a large portion of these would have been negative if the original primary outcomes declared on the ClinicalTrials.gov registry entry had been maintained and reported as originally planned, and if significant outcomes had not been promoted in place of non-significant outcomes. Furthermore, it is important to note that just because a study was published with negative results, does not mean that it was free of bias or spin. For example, a study with a high level of spin could be ultimately negative, but the results have been spun to seem more positive than initially intended.

With respect to measurable impacts after publication, positive studies were published more than one year earlier than negative studies (53 vs 61 months), again providing explicit evidence of greater propensity to publish positive studies. Whether this is due to investigators' greater motivation to submit positive results, or editors' propensity to prioritize publishing positive results, or both, remains uncertain. However, negative studies were associated with a significantly higher journal impact factor, and non-significantly higher citation scores. Therefore, although there is some evidence that positive studies are favoured post-publication, they may not be consistently advantaged in all aspects of attention scores.

Public efforts have been made in an attempt to increase trial registration to increase transparency in published research, and recent improvements have been reported.[10,12] However, researchers still seem to find non-significant results unsatisfactory, overlooking the importance of their contribution to the evidence base. A 2011 survey of authors indicated common reasons why outcomes were not reported included: failure to understanding why it is important to report negative results, fear of

data construed as uninteresting, not enough events worth reporting, and constraints imposed by the journal calling for brevity or more space.[31] Moving forward, focusing on supportive regulations and incentivized approaches to research that coincide with the issues at hand is vital in increasing transparency and integrity surrounding research.

3.4.3 Icon Array: Publication bias, Distortion, and Spin

As clearly shown by the icon array, there is a significant attrition in research from registration with original intentions, to eventual publication of results. Out of the sample of registered studies, only 34% were published, and of these, 61.6% were positive. However, only 36.2% of those reporting their originally intended primary outcome were positive. A large portion were 'turned positive' due to outcome distortion (13.5%) or spin (16.3%). In total, the proportion of registered studies that were published without outcomes being "turned positive" or without detectable spin was only 18%, and this represents a conservative estimate. If all aspects of outcome distortion were considered, the proportion of studies without outcome distortion or spin would be even less.

The cumulative effect of publication bias, outcome distortion, and spin is clearly shown by the over-representation of positive studies in the cohort of published studies. The impact is also seen in the propensity for positive studies to have higher levels of spin and distortion than negative studies. Furthermore, distortion and bias also exist in negative studies, where efforts did not go as far to manipulate outcomes or change the conclusion, though there was still some indication of attempts to mislead true results even in a sizable portion of negative studies.

3.4.4 Trends over Time

The results from the regressions offer insights into trends changing over time with respect to the proportion of positive results, and the proportion of results turning positive. A significant relationship was observed for the proportion of positive results and time, indicating significant increase in positivity over the years. In contrast, the trend was opposite for the proportion of studies that *turned* positive over time and was significant

for two exploratory analyses (studies that were turned positive due to spin, and studies that were turned positive overall). This might indicate that although positive studies are increasing over time, it may not be due to detectable distortion, insofar as distortion is detectable using the definition included in our study. Since distortion is difficult to trace, and since our definition of distortion does not claim to be a comprehensive definition, the estimates may be subject to detection bias and our limited definition. In addition, it may also indicate the improvement of clinical trial investigators in revising their stated primary outcomes on their ClinicalTrials.gov entry over time. Since we used the most recently declared primary outcome from the ClinicalTrials.gov RCT registrations, our definition of outcome distortion likely represents a conservative estimate compared to if we had disallowed investigators' revisions to primary outcomes definitions on the registry post-enrollment. Even so, concerns of publication bias and over-representation of positive results continue and should be further explored.

3.4.5 Strengths

This study should be interpreted in light of its strengths and limitations. Notable strengths include the large size of the cohort of RCTs examined, which is beyond the size of analogous studies in other areas of medicine that have evaluated outcome distortion and spin.[15, 16, 17, 18, 30, 31, 32] In addition, this study captured information across two decades in order to explore trends over time, which is beyond that attempted by most other analogous studies of questionable research practices. In addition to assessing distortion and spin, we also assessed evidence of the impacts of distortion and spin, including the proportion of studies that were positive, and the characteristics of studies associated with distortion, spin, and declaration of positive outcomes. We also conveyed the key outcomes of our study in an icon array, which communicates the attrition from study registration, through to enrolment, and publication, and further indicates the evidence of outcome distortion and spin (and the impact on positivity) within the cohort of studies that make it to publication. This innovative approach to data visualization will help to convey the complex interlinked concepts of publication bias, outcome distortion, and

spin, and may spur increased awareness of research waste and biased reporting on threats to objective, fully informed evidence-based decision making.

3.4.6 Limitations

A number of limitations existed when extracting information for trials registered in ClinicalTrials.gov. For example, we assumed that if enrollment size was above 0 in the registry entry, the trial was considered to have begun enrolment. However, this may not always have been the case, and it was beyond the scope of this study to determine the accuracy of enrolment sizes reported within ClinicalTrials.gov. In regard to the sampling strategy, it was designed to capture an evenly distributed sample of studies proportional the number of studies registered in each year and was large enough to view saturation with respect to proportion of studies with selective reporting or spin. This measure of saturation was chosen since it was the primary objective of the study. As a result, no efforts were made to ensure that this sample was proportional across other secondary or tertiary objectives, thus potentially bringing some disproportion to other aspects of the studies.

Similarly, when published studies were incomplete and unclear, a number of simplifying assumptions had to be made for primary outcomes, secondary outcomes, significance and direction of the conclusions. For example, when the primary outcome was not explicitly stated in the published study, the first outcome listed in the publication was taken as the primary outcome in order to compare reported versus pre-planned primary outcomes. Since we examined primary outcomes that were intended, at times they were not able to be fully classified as positive or negative since the originally planned primary outcome from the registered entry was not at all reported in the published study, requiring us to categorize the primary outcome positivity as “unknown”. Without access to unpublished data, it was not possible to determine the significance of these outcomes, and we assumed that they were “turned positive” if the reported outcome was positive (i.e., the outcome was switched from a non-reported primary outcome to a secondary positive outcome).

A further limitation is that our definitions for outcome distortion and spin were not mutually exclusive, and consequently, we could not report a composite estimate for this cohort of studies. In addition, our definitions for distortion and spin were likely conservative definitions and could be expanded in the future. Lastly, we also were conservative in our estimates of primary outcome switching, as we relied on the most recently reported primary outcomes reported in the RCT registry entry, rather than checking to see whether these had been changed post-hoc, or post-enrollment, by the investigators (as is allowable on ClinicalTrials.gov).

3.4.7 Conclusion

In conclusion, multiple approaches to manipulation of research in order to achieve positive outcomes is in evidence in the surgical evidence base. Over 61% of published studies were positive, many of which were “turned positive” through outcome distortion and spin. Positive studies had higher levels of detectable distortion and spin and were published faster than negative studies. This concern further escalates when considering that less than 34% of research that was planned and started has been published. To protect research integrity and reduce research waste, urgent action is required to ensure that evidence is published with full transparency, and that questionable research and reporting practices such as distortion and spin are thwarted.

Chapter 4

4 Categorizing and Describing Spin in HDI Categories

Conducting research proportionally to where the world's population resides is essential to improve applicability of research. In previous chapters, we have shown that most studies are conducted in countries with a very high and high Human Development Index (HDI). In this chapter, we further explore whether distortion and spin in studies conducted in very high and high HDI groups differed significantly from studies in medium and low HDI groups.

4.1 Introduction

Burden of disease and gross domestic product per capita are highly associated.[33] Multiple factors, such as vaccination access, quality healthcare, sanitation, nutrition, and housing, contribute to population health, social progress, and economic productivity.[34] As a result of multiple interposing factors, low-middle income countries (LMIC) experience a higher burden of disease compared to high-income countries (HIC). Access to safe, affordable, and timely surgery is scarce in low- and middle-income settings.[35] Additionally, lack of funding and resources reduces opportunities for LMIC to conduct the surgical randomized clinical trials, needed to generate evidence to inform safety and efficacy of interventions within context. Conducting high-quality research proportionally to where the world's population resides is an important consideration to improve generalizability and applicability of research. The objective of this study was to compare the characteristics (size, type of procedure, enrollment size) and extent of outcomes reporting distortion and spin in surgical RCTs registered between 1997 to 2017 for HIC versus LMIC settings.

4.2 Methodology

For this chapter, we used the same data set (n=1,718) as Chapter 3 (for full details of the methodology for extracting the data, refer to Chapter 3, Methodology, Section 3.2)

with the focus on comparing between HDI groups. Due to the small number of studies in low- and middle-income countries, the four HDI categories (low, middle, high, and very high) were simplified into two categories for our analysis: high income countries (HIC) (consisting of very high and high HDIs) and low- and middle-income countries (LMIC) (consisting of low and middle HDIs). Fifteen studies that did not provide information for where the study was conducted were excluded from the HIC versus LMIC analysis.

4.2.1 Statistical Analysis

Microsoft® Excel (v.16.51) was used to calculate summary statistics. Statistical inferences were calculated using RStudio (Version 1.2.5019).[28] This included the chi-square test for categorical data, and two-sample t-tests for continuous data (or Welch's two sample t-test when inequal variances were present). A two-sided p-value < 0.05 was considered significant.

Logistic regression models using a quasibinomial distribution were used to explore change over time in the proportion of positive studies, and the proportion of studies that turned positive over time, using similar methods as described in Chapter 3, with the addition of grouping the data according to HDI groups. Regressions were performed using RStudio (v.1.2.5019).[28]

4.3 Results

4.3.1 Included Studies

An outline of the search results leading to the 1,718 published studies included in this analysis, as well as a flow chart of studies, is provided in the previous chapter (Figure 6, Chapter 3). Study characteristics for the sample included in the analysis (n=1,718), and according to HIC and LMIC subgroup, are provided in Table 9.

4.3.2 Study Characteristics for High-Income Countries versus Low-Middle Income Countries

Study characteristics were reported for the entire sample of studies (n=1,718) and according to country income category: high-income countries (HIC) and low-middle-income countries (LMIC). Since not all studies were able to be categorized into a country, there are 15 studies (0.87%) that were missing from the HIC and LMIC characterization. Out of this total (n=1,703), 1,657 studies (97.3%) are from HICs, and 46 studies (2.7%) are from LMICs. Studies from LMICs took place in the geographical regions of the Eastern Mediterranean (21.6%), South-East Asia (54.9%), and Africa (23.5%); no LMIC studies from the Americas, Europe, or the Western Pacific. In contrast, most studies from HICs were in the Americas (28.2%), Europe (51.1), and some in the Western Pacific (12.9%).

Since our sample consists only of studies that had been published, it is expected that the status of these studies would be "Completed" on ClinicalTrials.gov. However, only 82% of registered studies from HICs and 91% of registered studies from LMICs were labelled as such, which indicates failure of some investigators to update their ClinicalTrials.gov registration upon trial completion. Completion status was similar between HIC and LMIC (p=0.28). Surgical categories differed significantly between HIC and LMIC (p=0.023). In HICs, the top 3 surgical categories were Orthopedic surgery (20.5%), Cardiac, Vascular and Thoracic surgery (14.7%) and then Gastroenterological surgery (13.9%), whereas in LMICs, the top 3 surgical categories were Gastroenterological surgery (26.1%), Orthopedic surgery (15.2%) and Obstetrics and Gynecological surgery (13.0%). On the other hand, in HICs, the bottom surgical categories were Breast (2.2%), Ophthalmic (2.2%) with several tied at 4.4%. In LMIC, there were no studies that examined Transplant surgeries, Plastic surgeries, Colon and Rectal surgeries, or Thyroid Surgeries.

The median enrollment sizes were comparable between HIC and LMIC (92 vs. 88), with a greater mean enrollment size in HIC, though not significantly different (327 vs. 205, p=0.058).

Table 8: Characteristics of Published Studies

Characteristic	Data Extracted Sample (n=1,718)			P-value
	Total (n=1718)	Country where study took place		
		HIC[1] (n=1,657)	LMIC[2] (n=46)	
	No. (%)	No. (%)	No. (%)	
Location by Region				
Americas	749 (27.7)	749 (28.2)	0 (0.0)	
Europe	1,356 (50.2)	1356 (51.1)	0 (0.0)	
Western Pacific	341 (12.6)	341 (12.9)	0 (0.0)	
Eastern Mediterranean	160 (5.9)	149 (5.6)	11 (21.6)	
South-East Asia	67 (2.5)	39 (1.5)	28 (54.9)	
Africa	30 (1.1)	18 (0.7)	12 (23.5)	
Surgical Category				
Breast	39 (2.3)	38 (2.3)	1 (2.2)	0.023
Orthopedic	352 (20.5)	340 (20.5)	7 (15.2)	
Neurology	27 (1.6)	25 (1.5)	2 (4.4)	
Transplant	70 (4.1)	69 (4.2)	0 (0.0)	
Obstetrics and Gynecology	178 (10.4)	171 (10.3)	6 (13.0)	
Plastic	33 (1.9)	32 (1.9)	0 (0.0)	
Urology	82 (4.8)	80 (4.8)	2 (4.4)	
Pediatric	82 (4.8)	80 (4.8)	2 (4.4)	
Cardiac, Vascular and Thoracic	247 (14.4)	244 (14.7)	3 (6.5)	
Otolaryngology	36 (2.1)	32 (1.9)	4 (8.7)	
Thyroid	19 (1.1)	19 (1.2)	0 (0.0)	

Dental, Oral and Maxillofacial	88 (5.1)	83 (5.0)	4 (8.7)	
Colon and Rectal Surgery	103 (6.0)	103 (6.2)	0 (0.0)	
Ophthalmic Surgery	67 (3.9)	64 (3.9)	1 (2.2)	
Gastroenterology	244 (14.2)	230 (13.9)	12 (26.1)	
General	51 (3.0)	47 (3.8)	2 (4.4)	
Completion Status[3]				
Completed	1,403 (81.7)	1350 (81.0)	41 (89.0)	0.28
Not Completed	250 (14.6)	242 (15.0)	5 (11.0)	
Ongoing	65 (3.8)	65 (4.0)	0 (0.0)	
Enrollment Size, median (IQR)	91.5 (51.0-186)	92 (51-190)	87.5 (51.0-150)	
Enrollment Size, mean (range)	321.6 (5-18,876)	327.0 (5-18,876)	204.6 (30-1,970)	0.058

1. HIC = High-Income Country
2. LMIC=Low/Middle-Income Country
3. Author reported from ClinicalTrials.gov registry

4.3.2.1 Outcome Distortion & Spin

Incidence of outcome distortion was 81% overall, and significant difference was not found for HIC versus LMIC studies (82% vs 74%; p=0.18). Sub-categories of type of outcome reporting distortion are outlined in Table 10. None of the sub-categories of outcome reporting distortion were significantly different between HIC and LMIC studies. Incidence of spin was 39% overall and did not differ significantly for HIC versus LMIC studies (39% vs 37%, p=0.75) for any sub-category.

4.3.2.2 Citation Bias & Journal Impact Factor

The proportion of positive studies did not differ significantly between HIC and LMIC studies (61.6% vs. 65.2%, p=0.62). Mean number of citations was significantly greater for HIC versus LMIC studies (75.4 vs. 14.7, p<0.00001). While mean JIF did not differ significantly between groups (8.9 vs. 5.2, p=0.24), there was significantly greater number of JIFs missing for LMIC studies (13.3% vs. 28.3%, p=0.0035). Mean time to publication did not differ significantly between groups (58.9 vs. 58.8 months, p=0.98).

Table 9: Statistical Comparison of High-Income Countries versus Low-Middle Income Countries

Characteristic	HIC[1] (n=1,657) No. (%)	LMIC[2] (n=46) No. (%)	P-value
Inadequately defined primary outcomes	237 (14.3)	4 (8.7)	0.543
Non-clinically relevant primary outcomes (reported)	175 (10.6)	6 (13.0)	0.80
Outcome distortion (at least 1 incidence)	1,353(81.7)	34(73.9)	0.18
Discrepancy in outcomes reported and outcome intended (primary and secondary)	1,329 (80.2)	34 (73.9)	0.16
Non-statistically significant primary outcome demoted	164 (9.9)	4 (8.7)	0.79

Statistically significant outcome promoted to primary	225 (13.6)	6 (13.0)	0.92
Intended primary outcome not reported at all	215 (13.0)	6 (13.0)	0.99
Primary outcomes were the same but with different time frames	185 (11.2)	5 (10.9)	0.95
Study "turned positive" due to primary outcomes yielding statistically significant results when intended had either non-statistically significant or unknown	225 (13.6)	5 (10.9)	0.26
Spin (any)	651 (39.3%)	17 (37.0%)	0.75
High Spin	121 (7.0)	2 (4.0)	0.45
Medium Spin	186 (11.0)	6 (13.0)	0.71
Low Spin	343 (21.0)	9 (20.0)	0.85
No Spin	1,006 (61.0)	29 (63.0)	0.75
Results "turned positive" due to spin in the conclusion or selective outcome reporting in the conclusion	271 (16.0)	7 (15.0)	0.84
Citation bias/Publication bias			
Number of Positive Studies Published	1,021 (61.6)	30 (65.2)	0.62

Number of Citations, median (IQR)	16 (5.0-50.0)	3.5 (0.3-14.0)	
Number of Citations, mean (range)	75.4 (0-3,024)	14.7 (0-188)	<0.00001
Journal Impact Factor, median (IQR)	3.5 (2.3-6.1)	2.2 (1.5-4.4)	
Journal Impact Factor, mean (range)	8.9 (0.17-91.2)	5.2 (0.09-79.3)	0.24
Journal Impact Factor, number not available	220 (13.3)	13 (28.3)	0.0035
Time to Publication (months) median (IQR)	50 (35.0-73.0)	48.5 (36.0-70.5)	
Time to Publication (months), mean (range)	58.92 (1-226)	58.8 (13-168)	0.98

1. HIC = High-Income Country
2. LMIC=Low/Middle-Income Country

4.3.3 Publication Rate

Table 10: Number of registered studies that made it to publication

	Sample (n=5,094)	HIC[1] (n=4,432)	LMIC[2] (n=90)	No Country Data (n=572)
	No. (%)	No. (%)	No. (%)	No. (%)
Stage of study				
Began Enrollment[3]	4,652 (91.3)	4,084 (92.1)	89 (98.9)	479 (83.7)
Published	1,718 (33.7)	1,657 (37.4)	46 (51.1)	15 (2.6)

1. HIC = High-Income Country
2. LMIC=Low/Middle-Income Country
3. Had an enrollment size larger than zero in its Clinical Trials.gov registry

An icon array was used to visualize the progression of studies from registration to publication, and to display the differences between HIC and LMIC in this progression (Figure 11). Out of the sample of 5,094 registered surgical RCTs, 4,652 (91.3%) of them began enrollment, and 1,718 (33.7%) were published. LMICs had the largest portion of studies making it to publication achieving a publication rate of 51.1% (46/90), and HICs had a publication rate of 37.4% (1,657/1,718). Studies with no country data provided had the smallest publication rate of 2.6% (15/572).

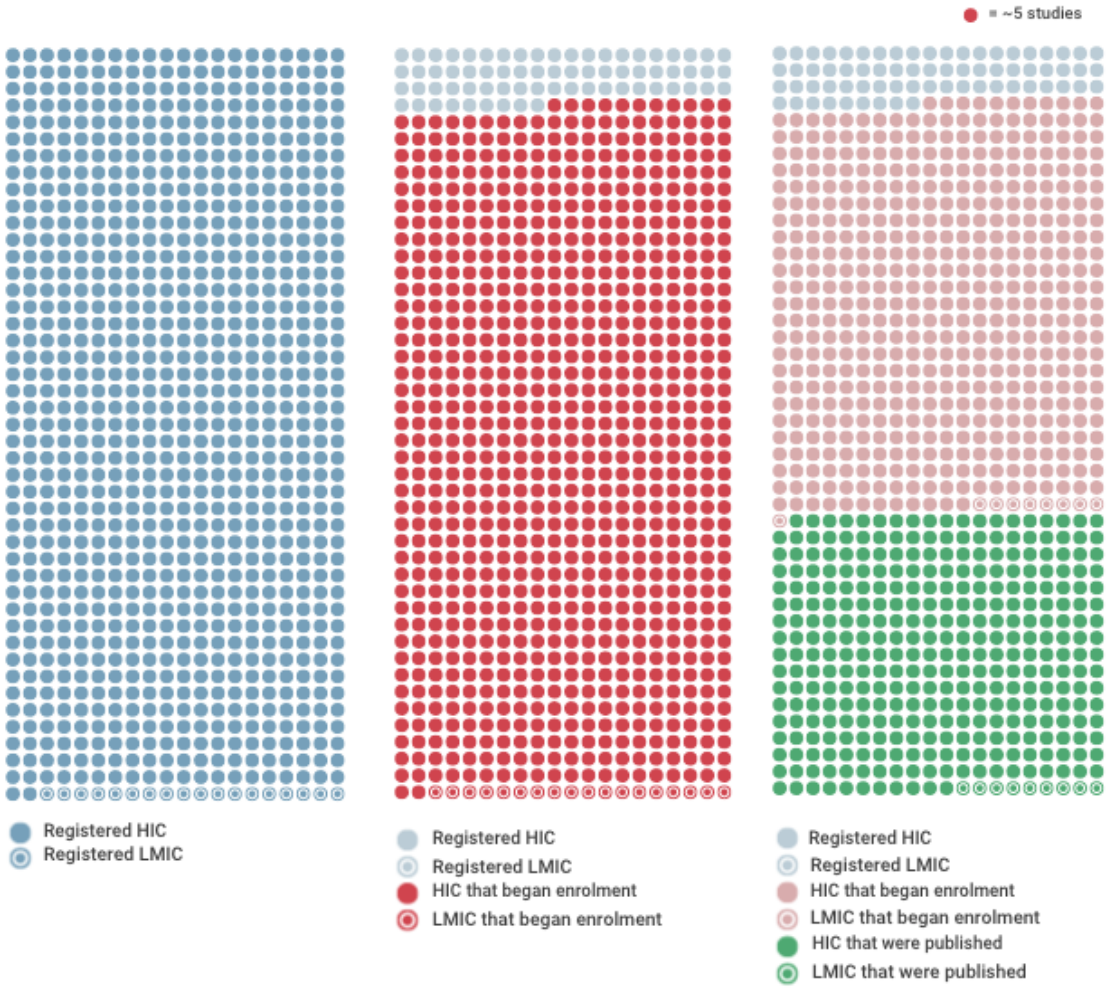


Figure 11: Studies analyzed from registration to publication separated into HIC and LMIC

Note: Studies without country data (n=572) were excluded from this figure, and thus there are 4,522 studies included for this analysis.

4.3.4 Production of Studies for HIC and LMIC

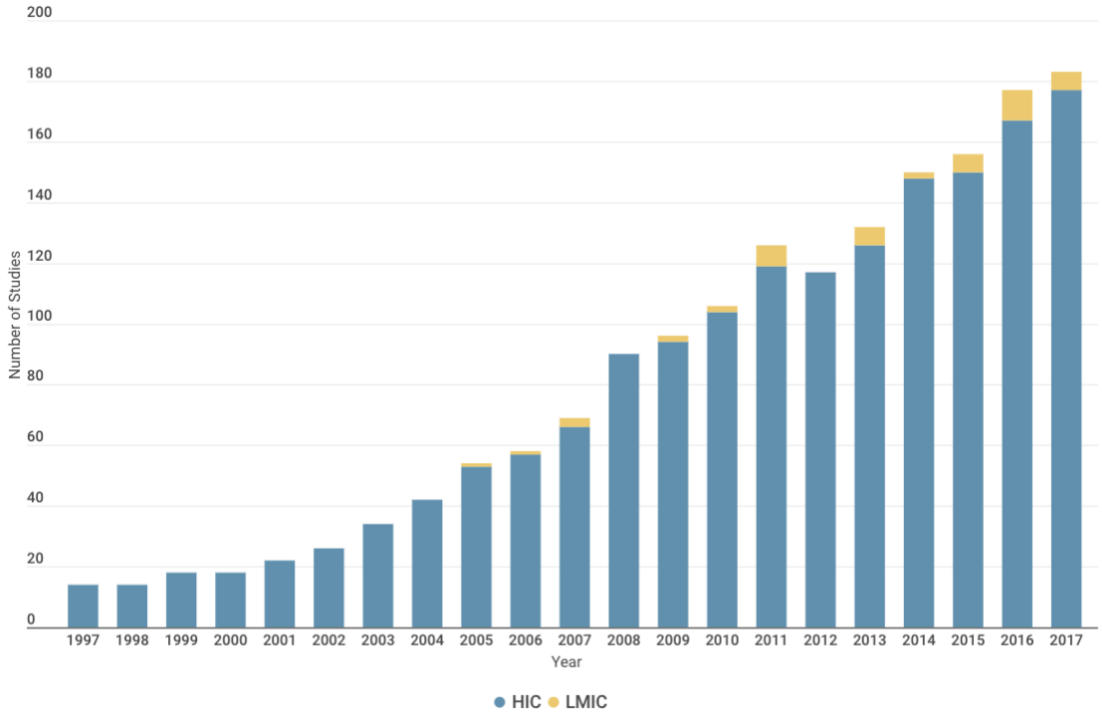


Figure 12: Stacked bar graph of number and proportion of HDI categories from 1997 to 2017

HICs had the greatest number of studies, with a growing trend across the years. LMICs experienced some growth over time, but still represent a minor proportion of studies (Figure 12). The proportion of studies from LMICs has not shown consistent growth. Registered surgical RCTs from LMICs did not appear in the ClinicalTrials.gov until 2005 and are absent in 2012.

4.3.5 Regression Results

Table 11: Logistic regression results for proportion of positive studies, proportion of studies turning positive out of all studies, and proportion of studies turning positive out of the positive studies subgrouped into high and low-middle income countries

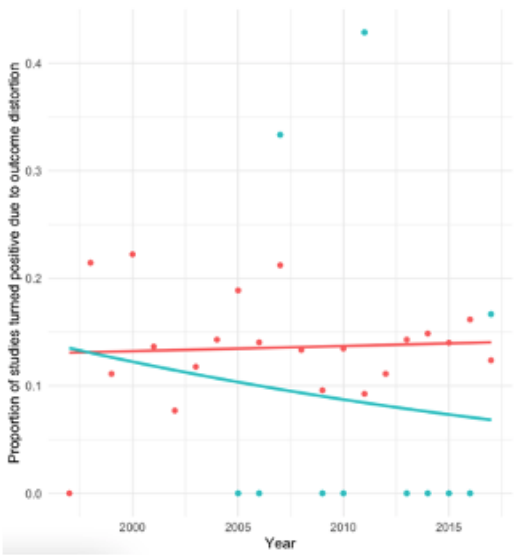
Proportion	Type of Distortion	Subgroup	Coefficient		
			Beta Estimate	Standard Error	p-value
Positive Studies	NA	HIC	0.040	0.013	0.0064
		LMIC	0.028	0.11	0.81
Studies that turned positive out of all studies	Due to outcome distortion	HIC	0.0041	0.016	0.80
		LMIC	-0.038	0.16	0.82
	Due to spin	HIC	-0.027	0.013	0.053
		LMIC	-0.023	0.16	0.89
	Overall	HIC	-0.0075	0.013	0.59
		LMIC	-0.075	0.16	0.66
Positive studies that turned positive out of positive studies	Due to outcome distortion	HIC	-0.020	0.020	0.32
		LMIC	-0.10	0.19	0.60
	Due to spin	HIC	-0.062	0.018	0.0028
		LMIC	-0.10	0.19	0.62
	Overall	HIC	-0.044	0.018	0.023
		LMIC	-0.16	0.21	0.46

1. Not applicable for the proportion of positive studies

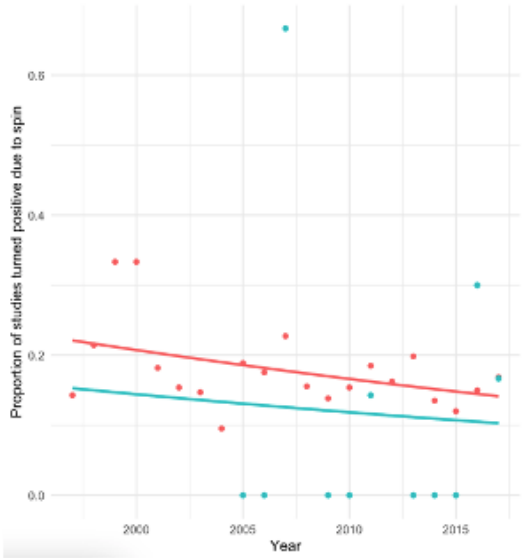
Logistic regression was attempted for the proportion of positive studies, the proportion of studies turning positive, and the proportion of positive studies turned positive grouped by HIC and LMIC. We qualitatively observed increasing trends in both HIC and LMIC for the proportion of positive studies produced each year seen (Figure 13), and this increase was significant only for HIC (Table 12: β=0.040, SE=0.013, p=0.0064).

Figure 13: Logistic regressions of the proportion of positive studies from 1997 to 2017 for HIC and LMIC (numerator is positive studies, denominator is all studies)

Figure 14 displays the proportion of studies that turned positive where we observed a greater slope of decreasing proportion over the years for LMIC (Outcome distortion p=0.82, Spin p=0.89, Overall p=0.66) than for HIC (Outcome distortion p=0.80, Spin p=0.053, Overall p=0.59); however, significance was not found.



a) Proportion of studies that turned positive due to outcome distortion



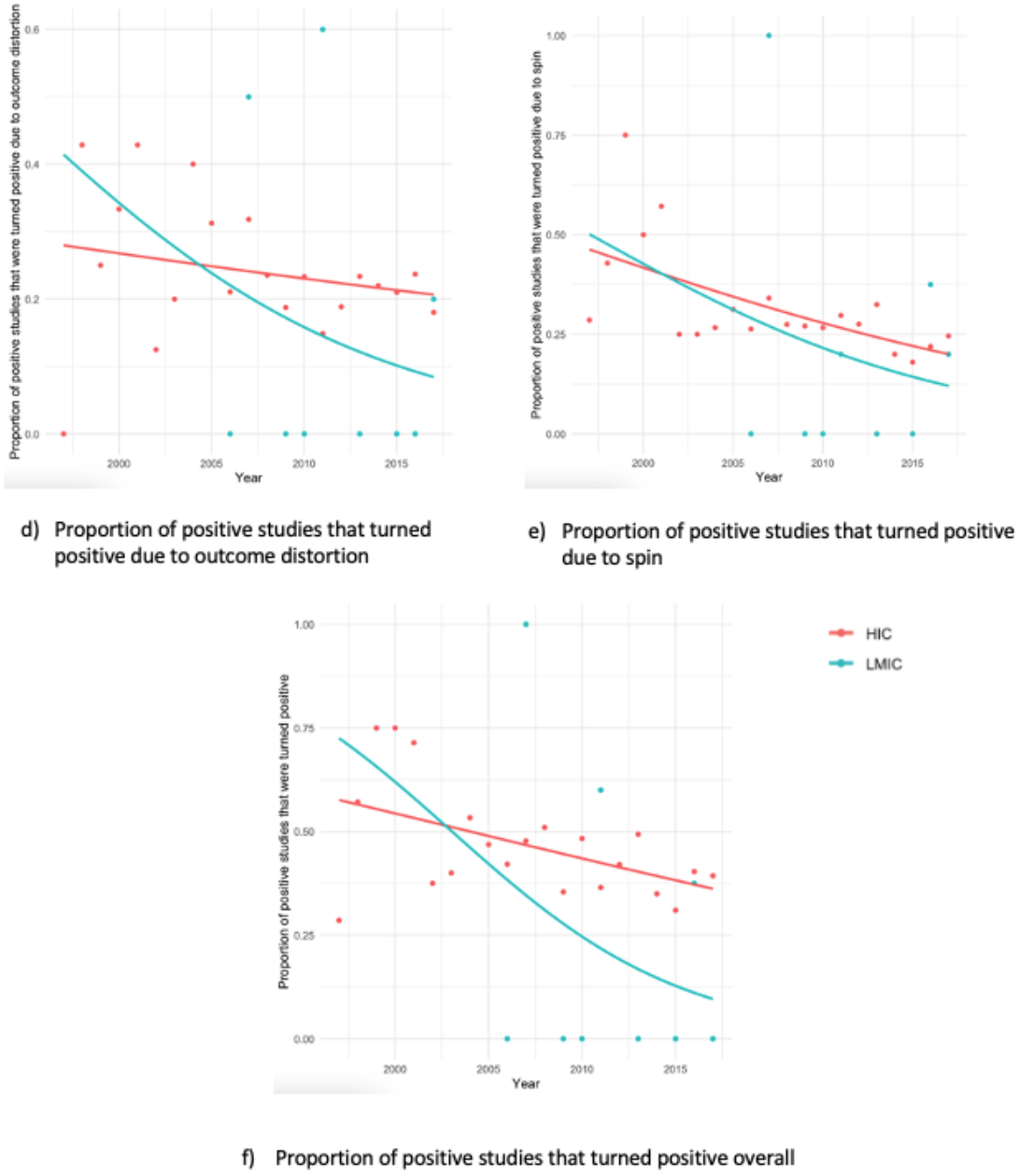
b) Proportion of studies that turned positive due to spin

c) Proportion of studies that turned positive overall

Figure 14: Logistic regressions of the proportion of studies that turned positive from 1997 to 2017 for HIC and LMIC (numerator is studies that turned positive, denominator is all studies)

As a sensitivity analysis, we also explored the proportion of positive studies that were turned positive, where the denominator is the number of positive studies (rather than a denominator of all studies) (Figure 15), and found a significant decline in proportion of

positive studies that were turned positive over time for HIC (Table 12 HIC for Spin β=-0.062, SE=0.010, p=0.0028 HIC overall β=-0.044, SE=0.018, p=0.023), with most of the decline related to proportion turned positive due to spin. Regressions for LMICs were underpowered due to few studies.



d) Proportion of positive studies that turned positive due to outcome distortion

e) Proportion of positive studies that turned positive due to spin

f) Proportion of positive studies that turned positive overall

Figure 15: Logistic regressions of the proportion of positive studies that turned positive from 1997 to 2017 for HIC and LMIC (numerator is studies that turned positive, denominator is all positive studies)

Overall, these regressions suggest that there is an increasing proportion of studies turning positive over the years; however, the number of studies that were *turned* positive is decreasing over time and is mainly attributable to decreases due to spin.

4.4 Discussion

4.4.1 Characteristics Compared to the World Population

Out of the cohort of studies included in our analysis of ClinicalTrials.gov between 1997-2017, an overwhelming proportion are from HIC (97.3%), which is not representative of the world's population distribution.[25] Only 2.7% of surgical RCTs were from LMICs, where 42.7% of the world's population resides (Table 13).

HICs show large and consistent growth in surgical RCT registrations over the years. In contrast, LMIC studies appear first in 2005, and show sporadic growth without stabilization in the upward trend. As a result, issues of generalizability to the world's population are seen and access and ability to conduct surgical research needs to be improved in the LMIC.

Table 12: Comparison of geographical location of studies compared to the world population. All values are displayed as percentages form the column total.

	Surgical Studies			**World Population 2021[25]**		
	Total	HIC (97.3)	LMIC (2.7)	Total	HIC (57.3)	LMIC (42.7)
Americas	27.7	28.2	0	13.2	21.8	1.6
Europe	50.2	51.1	0	12.1	20.6	0.6
Western Pacific	12.6	12.9	0	25.0	40.6	4.0
Eastern Mediterranean	5.9	5.6	21.6	9.3	6.4	13.2
South-East Asia	2.5	1.5	54.9	25.8	8.2	49.4
Africa	1.1	0.7	23.5	14.7	2.5	31.2

Most studies were from Europe (50.2%), Americas (27.7%), and Western Pacific (12.6%). In contrast, the world population has a much different order, where South-East Asia has the largest population (25.8%), followed by Western Pacific (25.0%). Thus, there is a large discrepancy between the proportion of RCTs performed geographically and the proportion of where the world's population lives which raises issues of research equity and generalizability.

4.4.2 Volume of Research & Growth over Time

Due to the costs, elaborate infrastructure, and education required for surgical procedures, access to surgery remains challenging for LMICs.[36] Therefore, it seems plausible that there are less surgeries occurring in LMIC, and fewer opportunities for surgical research. However, access to surgery is an essential component of universal healthcare and has been shown to be cost-effective for several life- and limb-saving procedures. Despite the known resource barriers and scarce access to surgical care, it is disappointing that only 2.7% of surgical studies have come from these areas, while 42.7% of the world's population lives in LMIC.[36] Together this highlights the need for improved resource distribution in these regions.

HICs show large and consistent growth in surgical RCT registrations over the years. In contrast, LMIC studies appear first in 2005, and show sporadic growth without stabilization in the upward trend. As a result, issues of generalizability to the world's population are seen as well as access and ability to conduct surgical research needs to be improved in the LMIC.

4.4.3 Characteristics Relating to Areas of Distortion

Categories of distortion were similar between HIC and LMIC. The number of positive studies and time to publication was similar between the groups. A standout difference was the greater number of citations for studies from HICs compared with LMICS. Cases for journal impact factor not found was higher for LMIC. When a JIF is not available it can suggest that the journal is not credible or does not meet criteria for

listing in the Clarivate JIF resource, thus potentially overestimating the JIF that was calculated for the LMIC and resulting in the true median and mean JIF to be even lower.[36]

4.4.4 Publication Bias within HIC vs. LMIC Studies

While many registered studies were started, just 33.7% made it to publication, which is direct evidence of publication bias. In our analysis, HICs had a higher non-publication rate than LMICs, also raising questions why HICs, with their greater access to resources and research, are less likely to publish than LMICs where resources and research are scarcer.

4.4.5 Regression Results

Regression analyses show that there is an increasing proportion of positive studies over time, qualitatively observed for both HIC and LMIC. The lack of power jeopardizes conclusions about difference in trends between groups over time. The higher proportion of positive results each year raises questions as to why more negative studies are not being published and is likely evidence of publication bias. The fact that the majority of studies are positive is suspect, given that the median size of the studies was low, and hence unlikely to be sufficiently powered to show positive effects most of the time. On the other hand, some might suggest that increasing proportions of positive conclusions also raises questions about whether more studies are positive due to more studies being conducted unnecessarily in the face of higher known a priori likelihood of success. However, this raises questions about whether an increasing number of studies are done where there is no longer equipoise, but rather the answer is already “known”, or a foregone conclusion. If this is the case, this would suggest possible research waste due to unnecessarily repeated studies for answers that were already known. Further research on the adequacy of study power and the presence of prior probability of this cohort of studies would be required to determine whether the increasing positivity in clinical trials is mostly due to spin, or whether there is also a propensity toward unethical or

unnecessary repeating studies with higher probability of success than equipoise would afford.

4.4.6 Limitations

This research should be interpreted in light of its limitations. Since country where the study has taken place was challenging to identify when not clearly listed in ClinicalTrials.gov, there may have been some misclassification of studies, though the risk is likely low since only 15 (0.87%) entries could not be categorized by country.

Another limitation includes the small number of studies in the LMIC category, resulting in underpowered analyses for subgroup comparisons of outcome distortion and spin. In particular, the logistic regressions may be unreliable due to the small number of studies from LMICs, with large dispersion.

In addition, the sample of studies included in this analysis was derived from registered RCTs with an identifiable and retrievable publication. Since publications from LMICs may be differentially identifiable and retrievable compared to HICs, this may have introduced bias in our dataset. Lastly, to account for studies that began enrollment, we made an assumption that studies with an enrollment size of larger than 0 the ClinicalTrials.gov registry indicated that the study had started, and some studies may have been missed due to failure of investigators to update their RCT registry page with enrolment numbers. However, this would be a concern only to the extent that registered trials without updates are systematically different than registered trials without updates on enrolment in the ClinicalTrials.gov database.

4.4.7 Conclusion

While the volume of surgical RCTs is increasing, more than 96.4% are conducted in HIC, where less than 57.3% of the world's population lives. Fewer than 2.7% of studies are conducted in LMICs, where the greater global burden of disease amenable to surgical care exists. This incongruity raises concerns about research equity, applicability, and generalizability. Our analysis did not provide evidence of differential risk of outcome

distortion and spin between HIC and LMICs. Nevertheless, HIC studies receive a greater number of citations, which correlates to greater opportunity for visibility and future funding, which may generate a self-fulfilling prophecy of continued incongruous research power in HIC settings. This research inequity should be addressed in future studies by journal editors, funders, policy makers, and research institutions in order to bridge the gap in LMIC surgical research.

Chapter 5

5 Integrated Discussion and General Conclusions

5.1 Overview

This research describes the characteristics of the current state of surgical research and identifies areas of distortion regarding research reporting. Randomized controlled trials (RCTs) registered on ClinicalTrials.gov from 1997 to 2017 involving patients undergoing a randomized allocation for a surgical intervention were extracted and analyzed to compare general characteristics, and relationships relating to distortion of results. These studies were further explored to focus on three overall objectives:

1. To describe the global body of surgical research for RCTs and identify disproportionate representation of specific characteristics including country and income-level where the study took place, surgical category, and completion status.

2. To determine the common areas of distortion in the surgical evidence base and quantify the overall level of distortion including failure to publish, spin and distortion of outcomes.

3. To identify and quantify the areas of distortion in the surgical evidence base and determine whether areas of distortion in the surgical evidence base is correlated to income-level of the country where the studies took place as we compared characteristics between high-income countries (HIC) and low-middle income countries (LMIC).

5.2 Integrated Discussion of Results

The function of evidence-based research serves to provide the highest quality of research for further implementation in healthcare and decisionmaking.[8] Thus, it is

necessary to ensure research is being produced with high integrity, that proportionally represents the world's population. Chapters 2 focuses on describing characteristics of the global surgical research pool to evaluate shortcomings in the global representation. Chapters 3 and 4 take a sample of the studies and explore the levels of failure to publish, as well as distortion and spin within the studies and comparisons of certain characteristics including differences between positive and negative studies, and differences between high-income and low-middle income countries respectively.

Chapter 2 showed an overall increase in the number of studies registered on ClinicalTrials.gov each year, creating a relatively large volume of surgical RCTs. However, a large portion of these studies remain unpublished, with this portion decreasing as the years progress. According to author-identified completion on the registry, only 57% of the overall studies reached completion with 84% completed in 1997 and 39% completed in 2017. Also decreasing throughout the timespan was enrollment sizes of studies, with a larger decrease in the completed studies. Additionally, a significant portion (76.8%) of studies have been conducted in very-high income countries, while only 19.9% of the world resides in these regions.

In Chapter 3, we aimed to conduct a more specific analysis of characteristics across studies and dive deeper into the studies to identify areas of spin or distortion. To do this, we obtained a sample of registries and searched them for the published studies. From the sample of 5,094, only 1,718 were published (34%). From this published sample, 62% declared a significant conclusion (1,058/1,718) of which 41% was not intended to be positive based on the registry, rather turned positive due to either spin or outcome distortion. Overall, the trend for reporting positive studies is significantly increasing over the years. However, the trend for studies that turned positive (out of the positive studies) is significantly decreasing over the years for turning positive due to spin and overall. Finally, positive studies are on average published in journals with lower impact factors compared to studies with negative conclusions (6.4 vs. 10.2, $p<0.00001$), while their medians are relatively similar (3.4 vs. 3.7), and there is no significant difference for number of citations between the groups ($p=0.26$). Positive studies are seen to have a

faster time to publication for mean (53.4 vs. 67.1 months, p<0.001) and median (46 vs. 58 months) and have more studies with unavailable journal impact factors (9.3 vs. 4.5, p=0.0051).

With a large proportion of studies produced in HIC, in Chapter 4 we aimed to see if income level of a country is related to differences in study characteristics and levels of spin and distortion. As seen in Chapter 3, a small proportion of studies made it to publication (34%). When comparing this between the income-levels, both groups had similar proportions that began enrollment (HIC: 92.1% vs. LMIC: 98.9%), while LMIC had a larger proportion that made it to publication (51%, 46/90) when compared to HIC where only 37% (1,657/4,432) made it to publication. This suggests that HIC have a greater failure to publish rate. There were no significant differences of spin or distortion between HIC and LMIC studies; however, HIC showed a significantly increasing trend of positive study production over the years (p=0.0064), while also showing a significantly decreasing trend of positive studies that turned positive over the years for turning positive due to spin (p=0.0028) and overall (p=0.023). While LMIC followed similar trends, none of them was significant. Finally, there were no significant differences between number of positive studies (HIC: 61.6 vs. LMIC: 65.2), journal impact factors (mean: 8.9 vs. 5.2 p=0.24, median: 3.5 vs. 2.2), or time to publication (mean: 58.9 vs. 58.8 months p=0.98, median: 50 vs. 48.5 months), between the groups. Significant differences were seen for the average number of citations larger in HIC (75.4 vs. 14.7, p<0.00001) with higher medians as well (16 vs. 3.5), and for journal impact factor not available, higher in LMIC (13.3% vs. 28.3%, p=0.0035).

5.3 General Conclusions

When conducting research, taking measures to achieve the highest quality of research is a necessary precaution in producing evidence-based research. Utilizing RCTs in order to achieve a higher level of research is exemplary; however, the quality needs to be integrated with a high level of integrity and careful attention when reporting the outcomes.[17]

Although positives are seen for the large volume of registered trials, concerns for publication bias and research integrity are raised as a result of significant unpublished work, and for the large proportion of increasing positive studies. Continued analysis of the unpublished work is necessary to make further conclusions.

Trial registration is intended to maintain transparency when reporting results, in order to reduce potential distortion that favour positive overall conclusions.[17] A relatively large portion of primary outcomes are inadequately reported, and lack of updates on the registry for the completion status suggests that the registries are not being adequately updated and carried through. High proportions of studies turned positive suggests that registry authors aren't being held accountable to their intended outcomes, and thus threatening the validity of the study. Although trends of decreasing studies turned positive is seen, further work for policy makers and those conducting research needs to be implemented to reduce this impact even further.

5.4 Future Directions

A large area of concern identified within this research was the large volume of unpublished work concerning registries that did not make it to publication. Future research should involve the exploration of these unpublished trials to explore how far they got and why publication has not occurred yet. This further investigation could uncover additional information into publication bias within this field and provide insights into why positive studies get published in a shorter time span, and why many of them are published in unknown journals. Reintroducing the unpublished studies could additionally strengthen the research base and aid in the evidence base practices for surgical research.

References

1. Sayek I. Surgery, surgeon, and measurement of value and quality of surgeons' work. Scandinavian Journal of Surgery. 2013 Sep;102(3):141-4.

2. Surgery. Cambridge Advanced Learner's Dictionary & Thesaurus ©. Cambridge: Cambridge University Press; 2021.

3. Gawande AA. Creating the educated surgeon in the 21st century. The American journal of surgery. 2001 Jun 1;181(6):551-6.

4. Weiser TG, Regenbogen SE, Thompson KD, Haynes AB, Lipsitz SR, Berry WR, Gawande AA. An estimation of the global volume of surgery: a modelling strategy based on available data. The Lancet. 2008 Jul 12;372(9633):139-44.

5. Safety WP, World Health Organization. The second global patient safety challenge: safe surgery saves lives. World Health Organization; 2008.

6. Meara JG, Leather AJ, Hagander L, Alkire BC, Alonso N, Ameh EA, Bickler SW, Conteh L, Dare AJ, Davies J, Mérisier ED. Global Surgery 2030: evidence and solutions for achieving health, welfare, and economic development. The lancet. 2015 Aug 8;386(9993):569-624.

7. Hariton E, Locascio JJ. Randomised controlled trials—the gold standard for effectiveness research. BJOG: an international journal of obstetrics and gynaecology. 2018 Dec;125(13):1716.

8. Evidence based medicine (EBM). *Treasure Island: StatPearls Publishing*.

9. Trends, Charts, and Maps - ClinicalTrials.gov [Internet]. Clinicaltrials.gov. 2021 [cited 15 August 2021]. Available from: https://clinicaltrials.gov/ct2/resources/trends

10. International Committee of Medical Journal Editors | About ICMJE | Clinical Trials Registration [Internet]. Icmje.org. 2021 [cited 15 August 2021]. Available from: http://www.icmje.org/about-icmje/faqs/clinical-trials-registration/.

11. ClinicalTrials.gov Background - ClinicalTrials.gov [Internet]. Clinicaltrials.gov. 2021 [cited 15 August 2021]. Available from: https://clinicaltrials.gov/ct2/about-site/background

12. FDAAA 801 and the Final Rule [Internet]. Clincaltrials.gov. 2021 [cited 15 August 2021]. Available from: https://clinicaltrials.gov/ct2/manage-recs/fdaaa

13. Rawat S, Meena S. Publish or perish: Where are we heading?. Journal of research in medical sciences: the official journal of Isfahan University of Medical Sciences. 2014 Feb;19(2):87.

14. Bornmann L, Mutz R. Growth rates of modern science: A bibliometric analysis based on the number of publications and cited references. Journal of the Association for Information Science and Technology. 2015 Nov;66(11):2215-22.

15. Boutron I, Dutton S, Ravaud P, Altman DG. Reporting and interpretation of randomized controlled trials with statistically nonsignificant results for primary outcomes. Jama. 2010 May 26;303(20):2058-64.

16. Khan MS, Lateef N, Siddiqi TJ, Rehman KA, Alnaimat S, Khan SU, Riaz H, Murad MH, Mandrola J, Doukky R, Krasuski RA. Level and prevalence of spin in published cardiovascular randomized clinical trial reports with statistically nonsignificant primary outcomes: a systematic review. JAMA network open. 2019 May 3;2(5):e192622-.

17. De Vries YA, Roest AM, de Jonge P, Cuijpers P, Munafò MR, Bastiaansen JA. The cumulative effect of reporting and citation biases on the apparent efficacy of treatments: the case of depression. Psychological medicine. 2018 Nov;48(15):2453-5.

18. Ioannidis JP, Munafo MR, Fusar-Poli P, Nosek BA, David SP. Publication and other reporting biases in cognitive sciences: detection, prevalence, and prevention. Trends in cognitive sciences. 2014 May 1;18(5):235-41.

19. McGauran N, Wieseler B, Kreis J, Schüler YB, Kölsch H, Kaiser T. Reporting bias in medical research-a narrative review. Trials. 2010 Dec;11(1):1-5.

20. Suñé P, Suñé JM, Montoro JB. Positive outcomes influence the rate and time to publication, but not the impact factor of publications of clinical trial results. PloS one. 2013 Jan 30;8(1):e54583.

21. Murad MH, Asi N, Alsawas M, Alahdab F. New evidence pyramid. BMJ Evidence-Based Medicine. 2016 Aug 1;21(4):125-7.

22. Research [Internet]. Knowledgetranslation.net. Available from: https://knowledgetranslation.net/knowledge-syntheses/ks-research/

23. Human Development Reports [Internet]. Hdr.undp.org. 2020. Available from: http://hdr.undp.org/en/2020-report

24. Regional offices [Internet]. World Health Organization. 2020. Available from: https://www.who.int/about/who-we-are/regional-offices

25. *Human Development Index (HDI) by Country*. World Population Review. 2020. Available from: https://worldpopulationreview.com/country-rankings/hdi-by-country.

26. Mlinarić A, Horvat M, Šupak Smolčić V. Dealing with the positive publication bias: Why you should really publish your negative results. Biochemia medica. 2017 Oct 15;27(3):447-52.

27. Journal Citation Reports Social Sciences Edition. Clarivate Analytics. 2021. Available from: https://clarivate.com/webofsciencegroup/solutions/journal-citation-reports/?gclid=CjwKCAiAtouOBhA6EiwA2nLKHx5R8yuFQdqQjEyVSa1-o1fG9XUYp1aV4qHH2fLL8Rjt_Bb9zHpQjhoC9XYQAvD_BwE

28. RStudio Team. RStudio: Integrated Development Environment for R. Boston, MA. (2019). Retrieved from http://www.rstudio.com/

29. Chalmers I, Glasziou P. Avoidable waste in the production and reporting of research evidence. The Lancet. 2009 Jul 4;374(9683):86-9.

30. Ross JS, Tse T, Zarin DA, Xu H, Zhou L, Krumholz HM. Publication of NIH funded trials registered in ClinicalTrials. gov: cross sectional analysis. Bmj. 2012 Jan 3;344.

31. Smyth RM, Kirkham JJ, Jacoby A, Altman DG, Gamble C, Williamson PR. Frequency and reasons for outcome reporting bias in clinical trials: interviews with trialists. Bmj. 2011 Jan 6;342.

32. Chow JT. Reducing Distortion–Identifying Areas to Improve the Quality of Randomized Clinical Trials Published in Anesthesiology Journals.

33. Kyu HH, Abate D, Abate KH, Abay SM, Abbafati C, Abbasi N, Abbastabar H, Abd-Allah F, Abdela J, Abdelalim A, Abdollahpour I. Global, regional, and national disability-adjusted life-years (DALYs) for 359 diseases and injuries and healthy life expectancy (HALE) for 195 countries and territories, 1990–2017: a systematic analysis for the Global Burden of Disease Study 2017. The Lancet. 2018 Nov 10;392(10159):1859-922.

34. Preston SH. The changing relation between mortality and level of economic development. Population studies. 1975 Jul 1;29(2):231-48.

35. Kolias AG, Rubiano AM, Figaji A, Servadei F, Hutchinson PJ. Traumatic brain injury: global collaboration for a global challenge. The Lancet Neurology. 2019 Feb 1;18(2):136-7.

36. Debas HT, Donkor P, Gawande A, Jamison DT, Kruk ME, Mock CN, editors. Disease Control Priorities, (Volume 1): Essential Surgery. World Bank Publications; 2015 Mar 23.

Appendices

Appendix 1: Search terms put into the ClinicalTrials.gov expert search

(randomised OR RCT OR randomly OR AREA[TitleSearch] randomized OR AREA[DesignAllocation] randomized) AND (AREA[InterventionSearch] (surgery OR transplant OR pancreatectomy OR pancreatoduodenectomy OR duodenectomy OR splenectomy OR nephrectomy OR lumpectomy OR mastectomy OR gastrectomy OR colectomy OR cholecystectomy OR appendectomy OR esophagectomy OR caesarean OR c-section OR hysterectomy OR oophorectomy OR thoracotomy OR arthroplasty OR “hip replacement” OR “knee replacement” OR prostatectomy OR rectopexy OR dissection OR bypass OR operative OR operatively OR laparotomy OR neurosurgery OR resection OR removal OR hepatectomy OR metastasectomy OR craniotomy OR “valve replacement”) OR AREA[TitleSearch] (surgery OR transplant OR pancreatectomy OR pancreatoduodenectomy OR duodenectomy OR splenectomy OR nephrectomy OR lumpectomy OR mastectomy OR gastrectomy OR colectomy OR cholecystectomy OR appendectomy OR esophagectomy OR caesarean OR c-section OR hysterectomy OR oophorectomy OR thoracotomy OR arthroplasty OR “hip replacement” OR “knee replacement” OR prostatectomy OR rectopexy OR dissection OR bypass OR operative OR operatively OR laparotomy OR neurosurgery OR resection OR removal OR hepatectomy OR metastasectomy OR craniotomy OR “valve replacement”) OR AREA[BriefSummary] (surgery OR transplant OR pancreatectomy OR pancreatoduodenectomy OR duodenectomy OR splenectomy OR nephrectomy OR lumpectomy OR mastectomy OR gastrectomy OR colectomy OR cholecystectomy OR appendectomy OR esophagectomy OR caesarean OR c-section OR hysterectomy OR oophorectomy OR thoracotomy OR arthroplasty OR “hip replacement” OR “knee replacement” OR prostatectomy OR rectopexy OR dissection OR bypass OR operative OR operatively OR laparotomy OR neurosurgery OR resection OR removal OR hepatectomy OR metastasectomy OR craniotomy OR “valve replacement”)) AND AREA[StudyType] EXPAND[Term] COVER[FullMatch] "Interventional" AND AREA[StartDate] EXPAND[Term] RANGE[01/01/1997, 12/31/2017]

Appendix 2: Results of search terms separated by year

Timeframe	Results		Timeframe	Results
01/01/1997 to 12/31/2017	24,740		01/01/2007 to 12/31/2007	1,027
01/01/1997 to 12/31/1997	62		01/01/2008 to 12/31/2008	1,338
01/01/1998 to 12/31/1998	104		01/01/2009 to 12/31/2009	1,390
01/01/1999 to 12/31/1999	125		01/01/2010 to 12/31/2010	1,542
01/01/2000 to 12/31/2000	163		01/01/2011 to 12/31/2011	1,706
01/01/2001 to 12/31/2001	227		01/01/2012 to 12/31/2012	1,751
01/01/2002 to 12/31/2002	298		01/01/2013 to 12/31/2013	2,035
01/01/2003 to 12/31/2003	373		01/01/2014 to 12/31/2014	2,238
01/01/2004 to 12/31/2004	569		01/01/2015 to 12/31/2015	2,573
01/01/2005 to 12/31/2005	722		01/01/2016 to 12/31/2016	2,755
01/01/2006 to 12/31/2006	861		01/01/2017 to 12/31/2017	2,881

Appendix 3: List of all countries represented along with their frequency, HDI value, HDI category and region according to the World Health Organization (WHO)

Country Name[1]	Frequency	HDI Value (UN) [2]	HDI Category	Region (WHO)[3]
Argentina	72	0.845	Very High	Americas
Armenia	6	0.776	High	Europe
Australia	200	0.944	Very High	Western Pacific
Austria	254	0.922	Very High	Europe
Azerbaijan	2	0.756	High	Europe
Bahrain	1	0.852	Very High	Eastern Mediterranean
Bangladesh	1	0.632	Medium	South-East Asia
Barbados	1	0.814	Very High	Americas
Belarus	13	0.823	Very High	Europe
Belgium	335	0.931	Very High	Europe

Benin	2	0.545	Low	Africa
Bosnia and Herzegovina	8	0.78	High	Europe
Botswana	1	0.735	High	Africa
Brazil	462	0.765	High	Americas
Brunei Darussalam	1	0.838	Very High	Western Pacific
Bulgaria	21	0.816	Very High	Europe
Burkina Faso	1	0.452	Low	Africa
Canada	1102	0.929	Very High	Americas
Chile	51	0.851	Very High	Americas
China	1052	0.761	High	Western Pacific
Colombia	46	0.767	High	Americas
Croatia	32	0.851	Very High	Europe
Cuba	1	0.783	High	Americas
Cyprus	2	0.887	Very High	Europe
Czechia	148	0.9	Very High	Europe
Denmark	488	0.94	Very High	Europe
Dominican Republic	3	0.756	High	Americas
Ecuador	1	0.759	High	Americas
Egypt	477	0.707	High	Eastern Mediterranean
Estonia	21	0.892	Very High	Europe
Ethiopia	5	0.485	Low	Africa
Finland	190	0.938	Very High	Europe
Former Serbia and Montenegro	2	0.8175[5]	Very High	Europe [4]
France	794	0.901	Very High	Europe
Gabon	1	0.703	High	Africa
Georgia	3	0.812	Very High	Europe
Germany	675	0.947	Very High	Europe
Ghana	3	0.611	Medium	Africa
Greece	128	0.888	Very High	Europe

Guam	2	0.926[5]	Very High	Western Pacific
Guatemala	2	0.663	Medium	Americas
Equatorial Guinea	1	0.477	Low	Africa
Haiti	1	0.51	Low	Americas
Honduras	1	0.634	Medium	Americas
Hong Kong, China (SAR)	59	0.949	Very High	Western Pacific
Hungary	102	0.854	Very High	Europe
Iceland	4	0.949	Very High	Europe
India	172	0.645	Medium	South-East Asia
Indonesia	20	0.718	High	South-East Asia
Iran, Islamic Republic of	80	0.783	High	Eastern Mediterranean
Iraq	9	0.674	Medium	Eastern Mediterranean
Ireland	77	0.955	Very High	Europe
Israel	246	0.919	Very High	Europe
Italy	577	0.892	Very High	Europe
Jamaica	1	0.734	High	Americas
Japan	127	0.919	Very High	Western Pacific
Jordan	7	0.729	High	Eastern Mediterranean
Kazakhstan	1	0.825	Very High	Europe
Kenya	7	0.601	Medium	Africa
Korea, Democratic People’s Republic of	693	0.916	Very High	Western Pacific
Kuwait	3	0.806	Very High	Eastern Mediterranean
Latvia	21	0.866	Very High	Europe
Lebanon	21	0.744	High	Eastern Mediterranean
Lithuania	35	0.882	Very High	Europe
Malawi	4	0.483	Low	Africa
Malaysia	50	0.81	Very High	Western Pacific
Malta	1	0.895	Very High	Europe

Mauritania	1	0.546	Low	Africa
Mexico	103	0.779	High	Americas
Moldova, Republic of	1	0.75	High	Europe
Montenegro	3	0.829	Very High	Europe
Morocco	4	0.686	Medium	Eastern Mediterranean
Mozambique	1	0.456	Low	Africa
Nepal	12	0.602	Medium	South-East Asia
Netherlands	416	0.944	Very High	Europe
New Zealand	68	0.931	Very High	Western Pacific
Niger	9	0.394	Low	Africa
Nigeria	8	0.539	Low	Africa
North Macedonia	1	0.774	High	Europe
Norway	257	0.957	Very High	Europe
Oman	51	0.813	Very High	Eastern Mediterranean
Pakistan	29	0.557	Medium	Eastern Mediterranean
Panama	8	0.815	Very High	Americas
Paraguay	1	0.728	High	Americas
Peru	20	0.777	High	Americas
Philippines	19	0.718	High	Western Pacific
Poland	220	0.88	Very High	Europe
Portugal	60	0.864	Very High	Europe
Puerto Rico	29	0.926[5]	Very High	Americas
Qatar	2	0.848	Very High	Eastern Mediterranean
Réunion	1	0.901[5]	Very High	Africa[4]
Romania	50	0.828	Very High	Europe
Russian Federation	151	0.824	Very High	Europe
Saudi Arabia	40	0.854	Very High	Eastern Mediterranean
Senegal	1	0.512	Low	Africa
Serbia	35	0.806	Very High	Europe
Sierra Leone	1	0.452	Low	Africa

Singapore	79	0.938	Very High	Western Pacific
Slovakia	20	0.86	Very High	Europe
Slovenia	27	0.917	Very High	Europe
South Africa	61	0.709	High	Africa
Spain	484	0.904	Very High	Europe
Sri Lanka	2	0.782	High	South-East Asia
Sudan	1	0.51	Low	Eastern Mediterranean
Sweden	395	0.945	Very High	Europe
Switzerland	295	0.955	Very High	Europe
Syrian Arab Republic	7	0.567	Medium	Eastern Mediterranean
Taiwan	189	0.761[5]	High	Western Pacific [4]
Tanzania, United Republic of	1	0.529	Low	Africa
Thailand	155	0.777	High	South-East Asia
Congo, The Democratic Republic of the	1	0.48	Low	Africa
Trinidad and Tobago	1	0.796	High	Americas
Tunisia	16	0.74	High	Eastern Mediterranean
Turkey	327	0.82	Very High	Europe
Uganda	15	0.544	Low	Africa
Ukraine	36	0.779	High	Europe
United Arab Emirates	5	0.89	Very High	Eastern Mediterranean
United Kingdom	604	0.932	Very High	Europe
United States	4022	0.926	Very High	Americas
Venezuela	3	0.711	High	Americas
Viet Nam	6	0.704	High	Western Pacific
Zambia	1	0.584	Medium	Africa
Zimbabwe	2	0.571	Medium	Africa

1. Country names were listed according to their United Nations 2019 name accompanying their HDI value. If not present, they were named according to their ClinicalTrials.gov reported name.
2. HDI values were obtained from the United Nations 2019 report.
3. HDI values were sorted into universally accepted HDI values as follows: Low below 0.550, Medium from 0.550 to 0.699, High from 0.700 to 0.799, and Very High from 0.800 to 1
4. Countries were assigned to regions according to the World Health Organization definitions of world regions and belonging countries.
5. Countries were not listed in any regions according to the World Health Organization, so they were placed in into regions based on their geographic location.
6. Countries that did not have HDIs available in the United Nations 2019 report were matched with countries that they most currently or most recently belonged to, or an average of various relevant countries. This includes Former Serbia and Montenegro using the HDI of the average of present-day Serbia and present-day Montenegro, Guam and Puerto Rico both being a territory of the United States using its HDI, Réunion using the HDI of France having held the status of a region in France and Taiwan using the HDI of the Republic of China.

Appendix 4: Surgical categories and the respective surgical procedures belonging to those categories along with the frequencies

Category	Types of surgeries	Frequency
Breast Surgery	breast (non-cosmetic), mastectomy, capsulectomy, lumpectomy,	380
Orthopedic Surgery	Arthroplasty, foot, hip, knee, wrist, hand, arthroplasty, arthroscopy, tibia, ankle, femoral, shoulder, orthopedic, elbow, spine, vertebroplasty, bunionectomy, Achilles, rotator cuff, osteotomy, joint, laminectomy, patellar, meniscectomy, acromioplasty, clavicle, Ewing, corticotomy, amputations, tenotomy, fasciotomy, tendon ruptures, laminoplasty, discectomy, curettage, osteosarcoma, myotomy, acromioplasty	2745
Neurology	neurology, nerve, pituitary, brain, neurosurgery, craniotomy, schwannoma, foraminotomy, craniectomy, posterior fossa, rhizotomy, neurectomy, epilepsy, carcinologic,	308
Transplant	transplant	738
Obstetrics and Gynecology	cesarean, cesarian, obgyn, hysterectomy, vaginectomy, oophorectomy, ovarian, myomectomy, uterus, episiotomy, sacrocolpopexy, endometriosis, endometrial abortion, large loop excision, cervical, episiotomy	1318
Plastic	contour, plastic, reconstruction, augmentation, reduction, cosmetic, panniculectomy, abdominoplasty, rhytidoplasty, lipectomy, mammoplasty, rhinoplasty, Mohs, burn, foot ulcer, scar, bichectomy	256

Urology	bladder, urinary, prostatectomy, kidney, nephrectomy, adrenalectomy, pelvic, Urethroplasty, sling, adrenalectomy, nephrolithotomy, prostate, vasectomy, Varicocelectomy	713
Pediatric Surgery	Children, Pediatric, infant, prenatal, neonate, neonatal,	482
Cardiac/Vascular/Thoracic Surgery	Cardiac, lung, chest wall, valve, vessel, vein, artery, heart, cabg, coronary, pulmonary, angioplasty, fontan, sternotomy, lobectomy, thoracic, atherectomy, esophagectomy, esophagus, thoracotomy, vascular, thromboendarterectomy, pleurectomy, lymphadenectomy, lymph node, Heller, sternum, sternotomy, rib, Mesothelioma, pleurectomy, bullectomy, segmentectomy	2198
Otolaryngology (ENT)	tympano, tonsillectomy, sinus, septoplasty, throat, nose, sinonasal, adenotonsillectomy, Laryngectomy, uvulopalatopharyngoplasty, ENT, cleft, neck, adenotomy, septorhinoplasty, dacryocystorhinostomy, parotidectomy, Laryngopharyngeal, vocal cord,	312
Thyroid	Thyroid, thyroidectomy	121
Dental, Oral and Maxillofacial Surgery	dental, molar, tooth, pulpotomy, gingivectomy, dental implant, periodontal, orthognathic, mandibulectomy, maxillofacial, pulpotomy, operculectomy, gingival	443
Colon and Rectal Surgery	hemorrhoidectomy, anoplasty, colectomy, rectal, colon, colorectal, colostomy, Rectopexy, ileostomy, polypectomy, haemorrhoidectomy, fistulectomy	754

Ophthalmic Surgery	Ophthalmic, cataract, eye, trabeculoplasty, opthalmic, vitrectomy, retinectomy, retinal, intraocular lens, canaloplasty, Keratectomy, trabeculectomy, pterygium, strabismus, blepharoplasty, iridectomy	583
General Surgery	any, general, splenectomy, appendix, umbilectomy, soft tissue, retroperitoneal, periampullary	471
Gastroenterology	Gastrectomy, cholecystectomy, gastrectomy, fundoplication, bariatric, hernia, abdominal, hernioplasty, gastric bypass, lichtenstein, digestive, stomach, GI, pancreaticoduodenectomy, pancreas, roux-en-y, splanchnicectomy, pancreatectomy, omentectomy, hepatectomy, liver, splanchnicectomy, hernioplasty, HCC, papillectomy, crohn	1939

Appendix 5: Details for Python program

A custom Python program was used to find the best-matched articles on PubMed and PubMed Central based on the available information provided in the registrations on ClinicalTrials.gov. This was done using 12 pieces of information from the individual ClinicalTrials.gov registries to be searched on the PubMed platform and returning a list of up to 3 of the most related published studies to the information from the registry. This was returned using study PMID or PMCID number. As a result, the corresponding articles from PMID or PMCID numbers were examined and compared against the ClinicalTrials.gov registry to see if they are matching and can be considered the published version of that registry.

Appendix 6: Definitions of spin for assessing conclusion spin according to Boutron et al.[15]

Level of Spin	Example	Explanation
None	"PCI [percutaneous coronary intervention] did not reduce the occurrence of death, reinfarction, or heart failure, and there was a trend toward excess reinfarction during 4 years of follow-up in stable patients with occlusion of the infarct-related artery 3 to 28 days after myocardial infarction."[13]	No spin: The conclusion is consistent with the results and highlights adverse events.
Low (acknowledgement of the statistically nonsignificant results for the primary outcome OR uncertainty and recommendations for further trials)	"A 50 mg of sublingual misoprostol 4 hourly for labour induction at term seems to have similar efficacy as 25 mg of vaginal misoprostol. Further studies on safety with larger numbers of women need to be conducted before routine sublingual misoprostol use in this setting."[14]	Spin: Conclusion of equivalence or similar efficacy Uncertainty in the framing: "Seems to have similar efficacy" Recommendation for further trials: "Further studies on safety with larger numbers of women need to be conducted"
Moderate (no acknowledgement of the statistically nonsignificant results for the primary outcome AND uncertainty or recommendations for further trials)	"Our data suggest that a structured nutritional intervention–physical exercise program is more efficacious than a nutritional intervention program in the reduction of global cardiovascular risk and cardiovascular risk factors, in only 16 weeks."[15]	Spin: Conclusion claims efficacy Some uncertainty in the framing: "suggest that"
High (no acknowledgement of the statistically nonsignificant results for the primary outcome AND no uncertainty AND no recommendations for further trials	"Combining ODS with colposcopy provides a clinically meaningful increase in the detection of CIN 2,3 in women referred for the evaluation of mildly abnormal cytology results."[11]	Spin: Focus only on statistically significant subgroup analyses. No consideration of the primary outcome on the whole sample. No uncertainty in the framing AND no recommendation for further trials

Appendix 7: Definitions and categorization of data extraction items

Category of items extracted	Item extracted	Definition and Source	Extraction Notes
Other relevant information	Clinical Trials URL	The study related registration URL link on clinicaltrials.gov	Extracted from RCT[a] registration pages on Clinicaltrials.gov
Other relevant information	Published article URL	The URL where the published article was found	Extracted from RCT registration pages on Clinicaltrials.gov

Other relevant information	Title of Clinical Trials registration	Title of the registered RCT	Extracted from RCT registration pages on Clinicaltrials.gov
Characteristic	Status of the registry	Whether the study is still enrolling, has been terminated, withdrawn, completed, etc.	Extracted from RCT registration pages on Clinicaltrials.gov
Characteristic	Condensed definition of study status of the registered RCT	Condensed version of the status of the registry in order to easily classify. Completed status on clinicaltrials.gov was categorized as 'completed'. Status of suspended, terminated, withdrawn, or unknown was reclassified as 'not completed'. Status of recruiting, not yet recruiting, active but not recruiting, ongoing, or enrolling by invitation was reclassified as 'ongoing'	Re-categorizes the given status of the registry into our simplified list of categories: completed, not completed or on-going
Characteristic	Country	Country where the study was conducted	Extracted from RCT registration pages on Clinicaltrials.gov. If not found, then extracted from any

			study-related publication(s).
Characteristic	HDI[b] number	HDI number associated with the country where the study was taking place.	Matched with listed countries using United Nations 2019 report.[23] If more than one country was present, the average was taken.
Characteristic	HDI category	Low below 0.550, medium from 0.550 to 0.699, high from 0.700 to 0.799, and very high from 0.800 to 1	Matched through HDI number.
Other relevant information	Condition/disease	The disease, disorder, syndrome, illness, or injury that is being studied.[c]	Extracted from RCT registration pages on Clinicaltrials.gov
Other relevant information	Interventions/treatment	A process of action that is the focus of a clinical study, typically a surgical procedure +/- other intervention for registered included RCTs.[c]	Extracted from RCT registration pages on Clinicaltrials.gov
Other relevant information	Arm	A group of participants in a clinical trial that receives a specific intervention/treatment or no intervention according to the trial's protocol.[c]	Extracted from RCT registration pages on Clinicaltrials.gov
Other relevant information	Surgical specialty	The category of surgery that the type of surgery in the study belongs to. This was	Extracted from RCT registration pages on

		divided into 16 categories known as breast, orthopedic, neurology, transplant, obstetrics and gynecology, plastic, urology, pediatric, cardiac, vascular, and thoracic, otolaryngology, thyroid, dental, oral and maxillofacial, colon and rectal, ophthalmic, gastroenterology and general.	Clinicaltrials.gov, matched with one of these 16 categories after analyzing the condition, interventions and arms.
Necessary information for further decisions	Primary outcome intended (POI [d])	The primary outcomes that were intended to be reported on, based on the RCT registration. This was repeated for the first 5 POI.	Extracted from RCT registration pages on Clinicaltrials.gov
Necessary information for further decisions	POI positive or negative	If the POI was significant it was labelled “positive”, and non-significant was labelled ”negative”. Alternatively, if significance was not clear it was assumed based on the language used. This was repeated for the first 5 POI.	Obtained from publication
Necessary information for further decisions	POI effect measure	The effect measure and outcome used to measure the POI was listed. This was repeated for the first 5 POI.	Obtained from publication
Necessary information	POI p-value/significance	Preferably the p-value was listed here to suggest the significance. If it was not	Obtained from publication

for further decisions		available, the method used to address significance was listed such as confidence intervals, language or indication of significance. This was repeated for the first 5 POI.	
Necessary information for further decisions	Primary outcome reported (POR[e])	The primary outcomes that the study actually reported on. If the primary outcomes (PO[f]) were not explicitly stated in the registry, the PO were extracted from the publication of the study. This was repeated for the first 5 POI.	Obtained from publication for up to 5 primary outcomes. If POI was not explicitly stated, the first results-based outcomes listed
Necessary information for further decisions	POR positive or negative	If the POR was significant it was labelled “positive”, and non-significant was labelled ”negative”. Alternatively, if significance was not clear it was assumed based on the language used. This was repeated for the first 5 POI.	Obtained from publication
Necessary information for further decisions	POR effect measure	The effect measure and outcome used to measure the POR was listed. This was repeated for the first 5 POI.	Obtained from publication

Necessary information for further decisions	POR p-value/significance	Preferably the p-value was listed here to suggest the significance. If it was not available, the method used to address significance was listed such as confidence intervals, language or indication of significance. This was repeated for the first 5 POI.	Obtained from publication
Necessary information for further decisions	Secondary/Other outcomes Intended	The secondary outcomes that were intended to be reported on, based on the registry.	Extracted from RCT registration pages on Clinicaltrials.gov
Necessary information for further decisions	Secondary/Other outcomes Reported	The secondary outcomes that were actually reported on. Only the outcomes were obtained, without mention of significance.	Obtained from publication
Distortion category	Are POI adequately defined?	Answered "yes" if the outcome has an unequivocal description with a method of measurement (if applicable), and a specified time frame.[6]	Reported after analyzing the POI from the registry
Distortion category	Are POR clinically relevant?	Answered "yes" if the outcome was relating to mortality, disability/functional status (e.g., ability to walk), disease/serious morbidity or discomfort (e.g., quality of life, pain).	Reported after analyzing all of the POR

Distortion category	Were the outcomes the same as planned?	Answered “no” if any of the outcomes differed from the registry outcomes and were not the same as planned (primary or secondary).	Reported after analyzing the intended outcomes compared to the reported
Distortion category	Were non-significant POI demoted?	Answered “yes” if any intended non-significant POI were demoted from a primary outcome.	Reported after analyzing the PO intended and reported
Distortion category	Were POI not reported at all?	Answered “yes” if any POI were not at all present in the study.	Reported after analyzing the PO intended and reported
Distortion category	Were the PO the same, but with different time frames?	Answered “yes” if the intended and reported primary outcomes were the same but had different time frames.	Reported after analyzing the PO intended and reported
Distortion category	Were significant outcomes promoted to POR?	Answered “yes” if the significant outcomes were added to the POR, but they weren’t mentioned as a POI.	Reported after analyzing the PO intended and reported
Necessary information for further decisions	Overall POI significance?	Overall significance of POI was reported in 6 classes as either significant (all of the outcomes were significant), mostly significant (if more than half of the outcomes were significant), mixed (if exactly half of the outcomes	Extracted from trial publication

		were significant), mostly non-significant (more than half of the outcomes were non-significant), non-significant (all the outcomes were non-significant) or unknown if the outcomes were not reported at all.	
Necessary information for further decisions	Overall POR significance?	Overall significance of POR was reported in 6 classes as either significant (all of the outcomes were significant), mostly significant (if more than half of the outcomes were significant), mixed (if exactly half of the outcomes were significant), mostly non-significant (more than half of the outcomes were non-significant), non-significant (all the outcomes were non-significant) or unknown if the outcomes were not reported at all.	Extracted from the trial publication
Distortion category	Did the results "turn positive" due to outcome distortion?	Answered "yes" if the POI overall significance was more significant in the hierarchy than the POI overall significance. Also noted as "yes" was if the POI was unknown due to being omitted	Reported after analyzing the overall POI significance and the overall POR significance.

		from the study, and the POR was reported positive.	Extracted from the trial publication.
Necessary information for further decisions	Was the abstract conclusion positive or negative?	If the abstract conclusion claimed an overall difference between interventions, then it was considered positive.	Obtained from publication
Necessary information for further decisions	Quote from the abstract conclusion	Direct quote of the abstract conclusion.	Obtained from publication
Necessary information for further decisions	Was the discussion conclusion positive or negative?	If the discussion conclusion claimed an overall difference between interventions, then it was considered positive.	Obtained from publication
Necessary information for further decisions	Quote from the discussion conclusion	Direct quote of the discussion conclusion.	Obtained from publication
Distortion category	Was either conclusion different from the POR to make the overall result “turn positive”?	Answered “yes” if the abstract conclusion was more significant in the hierarchy than the POR conclusion.	Reported after analyzing the overall POR significance compared to the conclusion significance. Extracted from the trial publication.

Necessary information for further decisions	Was there spin in the conclusion?	Answered "yes" if there was any conclusion spin in either of the conclusions (see Appendix 3.1 for details).	Reported after analyzing the conclusion quotes based on selected criteria[15]
Necessary information for further decisions	Quote of spin in the conclusion	Direct quote of where the conclusion spin was seen.	Obtained from publication
Distortion category	Level of conclusion spin (0-3)	The level of the conclusion spin indication, decided by predefined criteria and separated into any, low, medium, high, or none.[10]	Reported after analyzing the conclusion quotes based on selected criteria from the literature.[15]
General Distortion category	Overall suspicion of distortion	Overall judgement for general suspicion of distortion within the publication labelled as low, medium, high, or none.	Reported subjectively after reading publication and analyzing the POI, POR and conclusions
Other relevant information	Comments	Any comments for justifications of any extracted points, or explanations that were thought necessary to include.	Reported after data extraction was complete if necessary
Characteristic	Enrollment size from Clinical Trials	The enrollment size that was obtained for the study. The most recent enrollment size was used.	Extracted from RCT registration pages on Clinicaltrials.gov

Characteristic	Actual enrollment size	The actual enrollment size reported in the publication that was used for the study. If multiple were offered, the number of subjects randomized was used.	Obtained from publication
Characteristic	Start date	The date on which the registry was first available on ClinicalTrials.gov.	Extracted from RCT registration pages on Clinicaltrials.gov
Characteristic	Date of publication	The date of which the publication was successfully published.	Obtained from publication
Characteristic - Impact	Time to publication (months)	The time between the date of registration to the date of publication for a study.	Calculated by subtracting start date of RCT registration on Clinicaltrials.gov from date of RCT publication reported on journal URL
Characteristic - Impact	Number of citations	The number of citations received by the study.	Obtained by searching the article in Google Scholar
Characteristic - Impact	Journal	The journal where the study was published.	Obtained from publication

Characteristic - Impact	Journal Impact Factor	Journal impact factor (JIF[g]) of the journal of publication in the year of publication. For studies published in 2021, the JIF for the year 2020 was used. If journals or the year of the journal was not available in the journal citation report, it was labelled "not available".	Obtained from a journal citation report (Clarivate).[27]
General Distortion category	Number of times distortion occurred	The number of times an incident of distortion occurred either in the outcomes or in the conclusion, labelled as "distortion category" in this table.	Calculated by summing each time a "category of distortion" was deemed unfavourable

a. RCT= Randomized controlled trial
b. HDI= Human development index
c. Definition taken from ClinicalTrials.gov
d. POI=Primary outcome intended
e. POR=Primary outcome reported
f. PO=Primary outcome
g. JIF=Journal Impact Factor

Curriculum Vitae

Name:	Andrea Mataruga
Post-secondary Education and Degrees:	University of Waterloo Waterloo, Ontario, Canada 2015-2019 B.Sc. Western University London, Ontario, Canada 2019-present M.Sc.
Honours and Awards:	Western Graduate Research Scholarship 2019-2021
Related Work Experience	Teaching Assistant University of Waterloo 2017-2019 Graduate Teaching Assistant Western University 2020-2021