# RIGOR AND REPRODUCIBILITY OF CANCER MEDICINE EVIDENCE

By

CHRISTIAN COLE WAYANT

Bachelor of Science in Biology
University of Oklahoma
Norman, OK
2015

Submitted to the Faculty of the
Graduate College of the
Oklahoma State University
in partial fulfillment of
the requirements for
the Degree of
DOCTOR OF PHILOSOPHY
May, 2021

RIGOR AND REPRODUCIBILITY OF CANCER MEDICINE EVIDENCE

Dissertation Approved:

Matt Vassar

Dissertation Adviser

Al Rouch

Bruce Benjamin

Mousumi Som

Bavette Miller

# ACKNOWLEDGEMENTS

To those who are responsible for the successful completion of this dissertation:

First, to my wife and son, Elizabeth and Ezra, without you both I am one-third myself. I am forever grateful for your patience, encouragement, adaptiveness, strength, and love these past years. When I have considered my future, I have done so with you all in mind. I will never forget or take for granted how restorative your affections are for me. I not only learned how to be a scientist and scholar these past few years, but how to be a husband and father. Ezra, I will most remember how we played and laughed and learned together. I will to “fly” you until my arms give out. You are silly, kind, warm, and happy. Elizabeth, I will always show you reciprocity and support your dreams and goals just as you have supported mine. You are so caring, empathetic, selfless, and patient – all things I am sometimes not. I love you both immensely.

Next, to Dr. Vassar, I am glad to say that I accomplished my goal of leading a first-year team of Minions. Every other accomplishment is at best, second-rate. You are the first mentor, besides my parents, who I felt believed in me from the start. Your belief in me is what I most remember, and what will shape my career going forward (besides learning the true definition of “impact”). You know me as a person better than most, and you know that I would not have accomplished what I did without your support. Thank you for identifying and unlocking reasons for me to believe in myself.

To Dr. Som, you gave me a template for what hard work looks like in clinical practice. As best I can tell, your mentorship strategy is to challenge those around you to never settle for good enough, but to realize that you have more to give to yourself and to others. Thank you for teaching me that those values can coexist with compassion and selflessness. I will always consider your clinical mentorship as foundational in how I approach patient care.

To Dr. Rouch, Benjamin, and Miller, I cannot express enough thanks to you all for serving me as members of my doctoral committee. It was an honor to know and work hard for all 3 of you as I navigated my degree. You are all consummate academics and role models for me as I transition to my medical career. You all helped me navigate grant funding, gave me nuanced feedback on my research, and taught me how to mentor others. It is an honor to have your names in this dissertation document.

Finally, to my mom, dad, and brothers. You have all changed alongside me in the time I spent earning my doctorate. I am incredibly lucky to have been witness to all of it. Mom, thank you for supporting me unconditionally. You understand me better than anyone and always surprise me with how far out of your own way you will go to care for me and my family. Dad, thank you for being the man I want to be (even if you think that is crazy). You are an amazing father and Pops to Ezra. I consider you more of a friend each year and I cannot put into words how happy that makes me. Chase, you are one of my best friends and I will always be here to take care of, and support, you. I am lucky to live vicariously through you and Sarah while you are stationed in Italy. I will always be thankful to have you as a brother. I look up to you, and learn from you, more than you probably know. I cannot wait for you and Ezra to get to know each other. LJ, thank you for being the best uncle to Ezra. He loves to play with you and I know that as you both get older, you will be friends as well as family.

Name: CHRISTIAN COLE WAYANT
Date of Degree: MAY, 2021
Title of Study: RIGOR AND REPRODUCIBILITY OF CANCER MEDICINE EVIDENCE
Major Field: BIOMEDICAL SCIENCES

Abstract: The burden of cancer in the United States and abroad is comprised of significant morbidity, mortality, and psychological or financial harms. There remains a concern that the influence of published research is not maximized because of bias, lack of reproducibility, and suboptimal transparency. This dissertation comprises 10 investigations of such shortcomings. As a result of these 10 studies we first found that oncology journal policies on reporting guidelines and trial registration could be improved to strengthen the transparency in published research. We found that key improvements to oncology interventions in trials could facilitate better translation of published results to daily clinical practice. An investigation of financial relationships between oncologist-authors of influential trials and pharmaceutical drug firms uncovered pervasive, large, often undisclosed conflicts of interest. In a cohort of published trials, we found that oncologist authors misrepresented or distorted their findings to highlight favorable findings, even if this meant downplaying patient-centered endpoint results. We evaluated the potential harm from the publication of interim trial reports before patient-centered endpoints have accrued the necessary events to be fully powered. We reviewed a broad cohort of drug advertisements and found that drug firms omitted endpoints that were unfavorable, potentially compromising the integrity of the drug's advertised efficacy. We found that noninferiority trials, which are increasingly important in oncology research, were poorly designed and used statistical practices which may compromise their robustness. We turned to systematic reviews, finding that one's ability to reproduce the results of oncology meta-analyses was compromised by incomplete reporting of basic patient data. We found a significant risk of bias in systematic reviews cited by prominent cancer practice guidelines were at risk of bias. We investigated prominent cancer practice guidelines and found that patient values and preferences were undervalued. Altogether, the results of these ten studies indicate that oncology research requires a number of major and minor improvements to maximize its ability to work fully for the patient's benefit.

# TABLE OF CONTENTS

Chapter Page

# LIST OF TABLES

Table Page

Table Page

# LIST OF FIGURES

Figure Page

# CHAPTER I

# INTRODUCTION

### *The Burden of Cancer*

Approximately one in two men and one in three women in the United States will develop some form of cancer in their lifetime[1]. Not all of these cancers will result in death, and some that may have resulted in death are able to be treated. Therefore, it is estimated that only one in five men and women will die as a result of cancer. Fortunately, the risk of death from cancer has gradually, but steadily fallen since the year 2000[2], which is likely the result of a myriad of changes to how cancer is treated, diagnosed, and prevented. Treatments for cancer are being approved at a rapid pace, with close to 50 new anticancer medications receiving Food and Drug Administration (FDA) approval each year[3]. Cancer is also being discovered more often, likely because of public health calls for increased cancer screening. Certain cancers, like thyroid and melanoma, which both have low mortality rates, affect the rate of cancer diagnosis without significantly affecting the rate cancer mortality, as documented by several recent studies[4,5]. Nonetheless, the successful treatment of cancer appears to be improving at a steady, consistent rate.

New problems in the burden of cancer are emerging as cancer treatments advance. In particular, a new toxicity, referred to as financial toxicity, has received recognition which is not the direct result of the cancer drug, nor is it listed in the FDA label. Financial toxicity refers to a myriad of issues related to the decreased quality of medical care related to financial hardship.[6] These issues may manifest as the avoidance of medical care, limits to affordable therapies, or administration of incomplete courses of therapy. Cancer therapy is especially prone to financial toxicity because of the high cost and long duration of therapy. Many new oncology drugs are priced in the hundreds of thousands — the result of increased precision and thus smaller pool of patients who will receive them[7]. In cancer, however, the direct cost of a drug is not the only thing contributing to the overall cost. Other costs including hospital and surgical services, imaging, and radiation therapy may be the most expensive to patients[8]. New approaches to drug selection in oncology have included estimates of a drug's efficacy to price ratio and increased calls to lower the direct costs of drugs that receive market approval[9]. A perfect approach has not been identified, and the financial burden of cancer is likely going to remain a point of discussion in decades to come.

Given the high degree of morbidity and mortality to patients with cancer, it seems clear that all efforts should be made to improve the treatment of cancer as quickly as possible. This not only includes robust funding and innovation in the treatment of cancer, but also improvements to the efficiency and rigor of research as it is being published. It is well known that all research, medical and otherwise, is prone to bias and imprecision[10]. It is fortunate that oncology is a field that is driven primarily by randomized controlled trials, which are often considered the most robust form of primary research[11]. Problems in trial design, translation of the findings to patient care, and reproducibility of the results are persistent issues which may constitute meaningful barriers to maximizing the benefit of research funding. These issues are central to this dissertation which comprises 10 studies that investigate how cancer medical research can be more rigorous and reproducible.

### ***Statement of the Problem***

Rigorous and reproducible medical research is a fundamental prerequisite to cancer treatments that improve survival and quality of life. Such evidence is vitally important given the significant morbidity and mortality that results from cancer every year in the United States[1]. While our understanding of cancer biology and treatment strategies has gradually improved, there are existing concerns about the quality of cancer medicine evidence[12–14]. These concerns are not only limited to the rationale and design of published research, but also apply to oncologists and drug manufacturers. To improve how cancer is prevented and treated moving forward, mechanisms to improve and maintain the quality of medical research must be implemented. A prudent starting point would be to focus on the three aspects of cancer medicine evidence that exert the greatest influence on patient care: clinical trials, systematic reviews (SRs), and clinical practice guidelines (CPGs).

### *Conceptual Framework for the Study*

The research studies described herein apply meta-research methodology. Meta-research, also known as "research on research", is a novel method of research that blends aspects of SR and observational methodology. Meta-research, as it is applied here, does not involved individual patients. Rather, it involves research articles, journals, practice guidelines, or drug advertisements. Meta-research allows one to investigate key questions related to how research studies are designed, conducted, reported, and shared[15]. The common goal of many meta-research studies is to improve how research results are translated to improve the quality of patient care. The specific procedures followed in this dissertation are discussed below, but in general, these procedures follow those of other meta-research studies: database search, article screening, data extraction, data analysis, and data reporting. In all cases, where feasible, the data from these studies has been made publicly available via the Open Science Framework (OSF), which is an online repository for researchers to deposit data, protocols, pre-prints (completed studies without peer review), and post-prints (completed studies after peer review). Such open data is consistent with best practices to encourage the reproducibility and rapid translation of research findings to clinical practice.

Included are 10 investigations of bias, reporting, and transparency in oncology research studies. Five of these studies will be dedicated to clinical trials, since clinical trials are the most important study designs in cancer medicine for changing clinical practice. All FDA approvals for novel therapies must be based on at least one clinical trial. Two studies will be dedicated to SRs and will explore to what extent these studies are reproducible and transparent. One study will be dedicated to studying oncology CPGs, which are summary documents of all available scientific research on a given topic that are meant to guide patient care decisions. Two studies will be dedicated non-peer reviewed sources of oncology evidence: one to oncology journals and one to oncology drug advertisements. These last two studies gauge the extent to which oncology journals implement policies to improve the quality of oncology research, and the extent to which oncology drug advertisements accurately portray research information to the public.

### *Research Questions*

Two research questions were devised, which were purposefully broad to allow for study in multiple study designs and to fill gaps in knowledge in cancer research.

1. To what extent is oncology cancer evidence reproducible and rigorous?
2. What are the consequences of irreproducible or biased research?

### *Definition of Common Terms*

Common terms that would otherwise be unknown to persons unfamiliar with cancer or medical research will be defined here. Other terms that are used in a single study or used sparingly in this dissertation will be defined where they are mentioned. All definitions are taken from the National Cancer Institute database of definitions[16], unless otherwise stated.

1. Clinical Trials
    a. *Accelerated drug approval (AA)*
        i. An official process that allows a new drug to be approved by the U.S. FDA before it has gone through all of the required levels of testing in humans. It is only used for drugs that treat serious or life-threatening diseases for which other treatments may not be available or may no longer be effective. A drug may be approved through the accelerated approval process if it has shown certain signs in clinical trials that it might be beneficial for patients, such as a shrinking tumor. Further testing of the drug is required after it has received accelerated approval and is on the market to confirm that it really works.
    b. *Clinical Endpoints and Outcomes*
        i. Overall Survival

1. The length of time from either the date of diagnosis or the start of treatment for a disease, such as cancer, that patients diagnosed with the disease are still alive.

ii. <u>Quality of life</u>

1. The overall enjoyment of life.

c. *Surrogate endpoints:*

i. In clinical trials, an indicator or sign used in place of another to tell if a treatment works.

1. <u>Progression-Free Survival</u>

a. The length of time during and after the treatment of a disease, such as cancer, that a patient lives with the disease but it does not get worse.

2. <u>Response Rate</u>

a. The percentage of patients whose cancer shrinks or disappears after treatment.

SR definitions are taken from the Cochrane Handbook for Systematic Reviews of Interventions[17], unless otherwise stated.

2. <u>Systematic Reviews</u>

*a. PICO Question*

i. A key aspect of an SR that includes the Population(s), Intervention(s), Comparator(s), and Outcome(s) that will be included. Helpful in structuring the SR and increasing the SR findings to clinical practice.

*b. Meta-analysis*

i. An overall statistic (together with its confidence interval) that summarizes the effectiveness of an experimental intervention compared with a comparator intervention

*c. Heterogeneity*

i. Any kind of variability among studies in a systematic review, including clinical, methodological, and statistical variability.

No unique definitions for CPGs are necessary and an overview of CPGs will be given in the Literature Review.

### *Procedures*

In general, each investigation in this dissertation will adhere to the following structure:

1. Database search: We search PubMed (which includes MEDLINE) over a pre-specified time, usually 3-5 years, for articles published in oncology medical journals. Journals are selected from Google Scholar metrics, wherein articles are ranked by h5-index, which is a measure of citation rates of published articles over 5 years. The Google Scholar h5-index, while not perfect, is a better measure of overall journal popularity and influence of public articles. It improves upon Impact Factor ratings because all published articles are included in the h5-index.
2. Article screening: All articles retrieved from the database search are screened by two investigators according to pre-specified inclusion and exclusion criteria. Each screener is masked to the other's decisions, meaning each worked independently without collaboration, to eliminate bias. After screening, the two investigators would reconcile differences and achieve consensus on the cohort of included articles. Rayyan, an online article screening platform, was used for all screening.
3. Data extraction: In similar fashion to article screening, two investigators would extract data from all included articles. Masking was maintained and discrepancies were reconciled after completion. We used Google Forms to extract data for most, if not all studies.
4. Statistical analysis: For the majority of these studies, measures of central tendency and proportions were used to describe the data. For these analyses, Google Sheets was used. If more advanced statistical analysis was required, Stata 13 or 15.1 were used to analyze data.

### *Significance of the Study*

It has been estimated that 75-90% of scientific research experiments are not reproducible, which results in tens of billions of dollars of funding in the United States alone that are potentially wasteful[18]. Cancer medicine is one of the  highest funded, rapidly changing fields of medicine. A commensurate number of research studies are published each year[19] and annual funding of cancer research is in the billions[20]. Whether or not improvements in patient care keep pace with the amount of research being published is contingent on multiple factors, not least of which is the quality of the research being published. This dissertation will study specific forms of bias that are common across medical research, as well as provide background information on why mitigating these forms of bias is important. In cancer, where patients are vulnerable and rely on novel and experimental therapies to save their lives, biased research design or outcomes could be fatal. In cancer medicine, a single clinical trial can change the landscape of patient care overnight. Clinical trial enrollment for cancer patients represents a social contract in which patients get early access to experimental therapies while trusting trial investigators to use their data for the common good to prevent future harms in other patients. If a clinical trial is not designed, reported, or interpreted in an unbiased manner, patient data is not used to its maximum potential and the social contract is breached. There is little room for error in clinical trials and all methods to understand and prevent bias in the future must be investigated.

# CHAPTER II

# REVIEW OF THE LITERATURE

## *Overview*

To begin, it is important to understand the background and importance of clinical trials, SRs, and CPGs. These types of research outputs are the subject of the majority of studies included in this dissertation. This overview will describe the basic function, design, importance, and goals of each study design. An overview of the importance of clinical trials, SRs, and CPGs to the advancement of cancer clinical care will be discussed. From there, a discussion of how bias may affect study results will be had. Within sections on clinical trials, SRs, and CPGs, illustrative examples of how cancer research can improve will be used to explain the importance of this dissertation which investigated the rigor and reproducibility of evidence as it relates to cancer medicine.

### Clinical Trials

Clinical trials are fundamental to an evidence-based approach to patient care because many interventions have small to moderate effects sizes which may be obscured by chance or external factors[21]. The common analogy for why clinical trials are needed revolves around the use of a parachute when jumping out of an airplane. The risk of death when jumping out of an airplane is approximately 100%, with only extremely rare case examples of survival[22]. When a parachute is used, the risk of death plummets to only 1.1 deaths in 100,000, or 0.0011%[23]. One does not need a clinical trial of a parachute when jumping out of an airplane because the effect size is very large and intuitive. However, there are no "parachute" interventions in medicine and a review of 80,000 medical practices found that only one medical intervention — extracorporeal membrane oxygenation in premature infants with respiratory distress — had a reliably large effect size on mortality, equal to an odds ratio of < 0.2[21]. The role of a clinical trial in a medical system where easily identified, large effect sizes are extremely rare is plain: equalize all external factors to isolate a small to moderate effect size of an intervention in a given population under certain conditions. Clinical trials are useful in identifying interventions which improve patient lives one small step at a time. Well-designed clinical trials will always be superior forms of primary clinical research because of their ability to mitigate noise and identify signals.

The major way in which clinical trials separate noise and signal is through randomization and blinding. Often, a trial enrolls patients according to pre-specified inclusion criteria. Thus, the initial cohort of patients recruited are more homogeneous compared to the population with the disease or condition at large. However, other, possibly unknown or unmeasurable, factors may still exist that may affect the results of a clinical trial. Randomization is an ideal method to control for these unknown or unmeasurable factors[24]. If done properly, randomization starts the trial off with groups that look alike in all known and unknown ways. Blinding is equally important to the rigor of a clinical trial.[25] Trial patients, investigators, or assessors may be blinded, and when all three are blinded, the trial is "triple-blind". If a group is blinded, it means they are unaware of the intervention being received or given. Knowledge of the intervention may affect patient response, behavior, and attitudes toward an intervention. Similarly, investigator knowledge of an intervention may lead to biased assessments. This was the case in a placebo-controlled trial of multiple sclerosis patients, where only the unblinded investigator assessments found a significant effect of the intervention[26]. Last, when trial assessors — a third-party group separated from the investigators that are most often used in situations where blinding is impossible (e.g., intravenous vs. oral medication) — are not blinded, they may assess trial outcomes differently[27]. It may be said that randomization establishes an unbiased foundation for the trial, while blinding is instrumental in maintaining that foundation until the trial ends.

The advantages that randomization and blinding provide for identifying small to moderate effect sizes interventions, and as a result clinical trials have boomed in popularity in the era of evidence-based medicine[28]. Randomization and blinding should be considered the bare minimum for a clinical trial to be considered rigorous, and myriad other factors may affect a trial outcome[10]. As clinical trial popularity has increased, so has our expectation for their methods and reporting. Failure to employ the highest-quality methods may lead to results that are larger or smaller than they truly are. That is to say, every intervention has either no effect, positive effect, or negative effect. What is being determined in a clinical trial is the presence, size, and direction of the effect. Since a clinical trial is limited by pragmatism (e.g., limitations to the number of enrolled patients) and cannot precisely ascertain the truth, every clinical trial must be seen as a snapshot in time of an intervention's effectiveness. The next steps in advancing clinical trials are to maximize their rigor and usefulness in clinical practice by proactively mitigating and identifying sources of bias. This will help ensure that every snapshot of an intervention is as accurate and truthful as possible. It has been understood that maximizing trial effectiveness requires 1) a mechanism to increase transparency; 2) an audit of trial methods and design; 3) the choice of patient-centered endpoints; and 4) reporting and disclosing all trial information.

### *Increasing Trial Transparency*

The concept of trial registration was borne out of efforts to maximize altruism and trust in clinical research.[29] The impetus for these efforts revolves around the fact that clinical research involves human participants who contribute their time and data in exchange for access to cutting-edge drugs and advancements in clinical care. Many in the scientific community see the inclusion of patients in clinical trials as a social contract: patients contribute their data with the understanding that their data will be used and reported in an ethical manner, regardless of trial results[30]. It is plainly understood that patients have a right to access and make decisions about their data, be it in a clinical trial or elsewhere, but individual patient data are often not available to the public or the patient participants[31]. Before trial registration, there was no way to track what clinical trials had been started around the world. Without a way to track which trials have been started, some trials that enrolled patients and that generated data may go undetected and unpublished. In 2005, to increase transparency, promote honesty, and combat conflicts of interest in clinical trials, the International Committee of Medical Journal Editors — the most widely recognized coalition of medical journals — declared prospective trial registration to be a requirement for publication moving forward.[29] Following that declaration, in 2007 the Food and Drug Administration Amendments Act (FDAAA) made clinical trial registration a legal requirement for any trial that enrolled patients in the United States and tested FDA-regulated interventions.[32]

Trial registries employ a number of useful practices to monitor trial progress and evolution. As of 2016, there were 17 documented trial registries, some named for countries that founded them (e.g., Australian New Zealand Clinical Trials Registry) and some deliberately named to recruit an international cohort of trials (e.g., International Standard Randomised Controlled Trial Number registry).[33] ClinicalTrials.gov (CT.gov) is the most popular registry, and the one most useful to discuss clinical trial features. CT.gov is a free, publicly-available website that houses time-stamped entries and changes to individual trial registrations. In CT.gov, studies are labeled as initiated, ongoing, completed, or unknown. Results may be posted when available. In addition, the time-stamps are attached to every line item, allowing the public to review a history of changes for everything from the title to the investigation sites to the endpoints being measured (and their order). The result is a database that complements official study protocols and provides a means of detecting where changes that result in bias may have entered a trial.

The two goals of trial registration, as outlined by the directors of CT.gov were to 1) "establish a publicly accessible and searchable database for disseminating a minimum set of structured information about all ongoing and completed trials", and 2) "provide access to date-stamped protocol details throughout the study lifecycle".[33] Overall, trial registration has been successful at accomplishing its two key goals, but not without some noticeable areas in need of improvement. With respect to goal one, CT.gov houses 358,767 individual trial registrations from all 50 states and 219 countries, as of November 2020.[34] One may search and filter by a number of pertinent clinical factors, including, but not limited to, disease or condition, intervention, eligibility criteria, and study locations. Thousands of research studies[33] have been conducted that use CT.gov, either for primary research to audit the completeness of registration reporting[35], detect bias in published trials[36], or for inclusion of unpublished data in an SR. The administrators of CT.gov self-identified several areas of improvement, including incomplete registrations, unidentified duplicate registrations across databases, out-of-date registrations (e.g., those without follow up from authors), and retroactive registration after trial start, which may obscure early changes in trial design and blind the public to possible bias[33].

Goal two includes access to the date of trial registration broadly, and the time-stamped primary outcome measure entries with sufficient detail to allow for detection of unacknowledged changes.[33] This goal is vitally important to prevent two major forms of bias which may distort the portfolio of published research: 1) publication bias and 2) selective outcome reporting bias. Publication bias refers to the selective publication of research studies that reach favorable conclusions, typically those that are statistically significant. Selective outcome reporting bias refers to the selective inclusion, exclusion, alteration, or reordering of study endpoints, often in the pursuit of conveying favorable study results.

### *Audits of Trial Methods and Design*

With the advent of the evidence-based medicine movement in the 1980s has come an emphasis to conduct meta-research. Meta-research is simple in its aims: to audit swaths of research to improve how we perform, communicate, verify, evaluate, and reward research[37]. Meta-research may encompass a wide range of observational, interventional, and theoretical designs. Some meta-research studies generate models to ask critical questions about the utility of popular research designs[38]. Some may ask whether the control arm in a clinical trial is considered standard of care[39], since a suboptimal control arm may not capture the true magnitude of benefit that a novel intervention adds to a field. Another meta-research study may ask whether conflicts of interests are disclosed according to rules and policies set forth by medical journals or CPG panels[40,41]. Last, a meta-research study may evaluate the risks of bias in a cohort of research studies with common features[42,43]. Despite the diversity of meta-research studies, they all share one common goal: optimizing scientific research.

There is a strong need for meta-research in cancer. Currently, oncology research is being conducted at a breakneck pace unlike anything seen in history. CPGs often require semi-annual updates given the number of practice-changing trials that are published[44]. However, there has been growing concern about whether the oncology community is able to self-police itself and reject research that is flawed or suboptimal[45]. It would seem that growing social pressure to cure cancer[46], paired with the daily loss of life in an oncologist's practice, and ample money from pharmaceutical companies in the form of direct payments to physicians[41,47] has led to a low-bar for Food and Drug Administration drug approval[3] and guideline recommendations[48]. Oncology clinical trials often have significant methodological shortcomings, including inadequate use of patient crossover (when patients in the control arm move to the intervention arm)[49], suboptimal control arms[39], poor statistical assumptions[50], misrepresenting and distorting research findings[51], and biased interpretation of research findings[52]. Meta-research studies are responsible for all of these findings and will continue to be important to understand how best to optimize oncology trials in the future.

### *Choosing Patient-Centered Endpoints and Outcomes*

In clinical research, the goal is not to simply conduct a study and derive an outcome — that outcome must mean something for patients. It is the goal of clinical research to test what is meaningful, not simply measurable[53]. Patient-centered outcomes (e.g., measured variables, like fatigue score) and endpoints (e.g., measured parameters, like change in fatigue score over 6 weeks) can be wide-ranging and are best identified using the help of patients and other stakeholders. The inclusion of patients into these decisions about patient-centered endpoints has been the mission of groups such as COMET (Core Outcome Measures in Effectiveness Trials)[54], PCORI (Patient-Centered Outcomes Research Institute)[55], and SPOR (Strategy for Patient-Oriented Research)[56]. Patient-centered endpoints may be pragmatic, like increased exercise capacity in a patient with heart failure, or could be absolute, like reduction in death due to heart failure. Regardless, all patient-centered endpoints have one thing in common: they matter to patients.

In oncology, there are really only two endpoints that are commonly measured that are patient-centered: Overall Survival (OS) and Quality of Life (QoL). The rest of commonly measured variables in oncology trials (all defined in the Introduction), such as progression-free survival (PFS), are surrogate endpoints and not wholly patient-centered. The main argument using surrogate endpoints is to save time in clinical trials, since often these endpoints can be measured before a patient dies or follows up to evaluate quality of life. This argument has been largely refuted, but persists nonetheless[57]. It has been suggested that slowing tumor growth, part of what is measured with PFS, is valuable to patients, but this has not been empirically proven. The few studies that asked patients whether PFS matters to them were synthesized recently and found to be significantly flawed, since they did not define PFS properly to patients[58]. PFS is a composite endpoint that measures the time to 1) growth of a known tumor by 20%, 2) development of new tumor lesions, 3) death of the patient − whichever happens first[59]. The first two measurements are measured using regular computed tomography scans in trial patients, with the third being assessed at regular intervals using patient follow up and contact. It can be theorized that growth of a tumor or development of new lesions is meaningful to patients, but this assumes a patent can feel their tumor grow. In many cases, the patient cannot tell when their tumor grows at all, which explains why lung cancer patients present with advanced disease so often[60]. In reality, the assessment of PFS is largely subjective, since tumors often grow before a patient dies (removing the possibility of death being the event of interest in PFS), and there is no evidence to suggest that 20% growth is what matters to patients. All of these factors notwithstanding, PFS is the most common endpoint that leads to FDA approval in oncology, signaling that pharmaceutical companies, trial authors, and regulators are content with its measurement[3].

The use of PFS and other surrogate endpoints for drug approval is not the only reason why they are not patient-centered endpoints. A reasonable person may see the flaws with PFS measurement, for example, and still see value in it if it predicts improvement in patient-centered outcomes, like OS. This hypothesis has been refuted time and time again. In an updated SR, it was found that surrogate endpoints do not predict OS[61,62] in the majority of cancers. Moreover, different SRs showed that surrogate endpoints do not predict improved QoL in cancer patients[63,64]. For oncology research to accurately predict patient outcomes in the real-world, improvements to trial design must be made, chief among them must include the measurement of patient-centered endpoints.

### *Reporting and Disclosing All Trial Information*

For clinical trials to maximize its effectiveness, all individual patient and aggregate trial-level data must be made publicly available. The reason for this is simple: the history of medical research has shown that conscious and unconscious bias in how data is sorted, analyzed, and presented may affect trial results and conclusions. Data sharing is now considered an ethical requirement for clinical trials by the ICMJE, similar to trial registration[65]. The ethical and pragmatic benefits of trial data sharing include improved accuracy of trial results[66], advancements in scientific discovery via secondary analyses[67], and accelerated scientific progress[68]. Beyond these commonly understood reasons, there is also a strong argument that the patients who contribute data to the trial are the true owners of the data, similar to how patients own their data in everyday clinical encounters[30].

Previous research has suggested that clinical trial authors are willing to share trial data publicly, but barriers still persist that may prevent complete adherence[69]. These barriers, for example, may include fear of scrutiny from third-parties who aim to re-analyze study results or lack of familiarity with how data must be formatted or how to use data repositories. Another significant barrier is that sharing one's data does not positively affect career advancement, but the uncovering of accidental (or intentional) flaws in a dataset may negatively affect career advancement. From the point of view of a trial author, absent external compulsions from a funding source or journal, there is little incentive to share one's data. Overcoming these barriers requires innovative approaches, such as incentive programs that counteract negative pressures to sharing data[70] or alterations to how academic advancement is achieved.

In oncology, the availability of data sets is complicated by the overbearing influence of the pharmaceutical industry in leading clinical trials. Pharmaceutical companies maintain proprietary control of the individual patient data that results from their clinical trials. The result is that data is only available upon request of the company. Despite endorsing a commitment to sharing anonymized individual patient data in 2014, the pharmaceutical industry has been found to rarely share data, due to the severe restrictions for which trial data may be shared[31]. A recent analysis of the availability of industry-sponsored trial data found that only 9/61 (15%) identified trials were eligible for data sharing two years after trial completion. The main reasons for data not being available were that the pharmaceutical company did not have a data sharing policy or process and that trials were still ongoing. These barriers signify that commitments to data sharing may be toothless if no consequences or incentives exist to reinforce them.

### *Summary*

Overall, clinical trials are the best primary research method to identify which interventions are effective and which are not. The backbone of a rigorous clinical trial is randomization and blinding. Additional measures must be taken to cement trials as unbiased, useful research studies, however. Without transparent declaration of one's methods before study initiation, bias may seep in and affect results. Without choosing a patient-centered endpoint, the results may not matter. Finally, without publication of independent patient data in a public repository, results may not be completely trustworthy and scientific advancement may be stalled.

## Systematic Reviews

SRs are studies that aim to "synthesize all empirical evidence that fits pre-specified eligibility criteria in order to answer a specific research question."[17] The key feature of SRs that distinguish them from other forms of research are that they use systematic research methods, which can be accomplished by combining multiple database searches with reviews of the unpublished literature. The benefit of SRs is the ability to derive summary effects by combining all retrieved research data, and from these summary effects, identify more robust answers to key clinical questions. For example, the popular SR of antenatal corticosteroids for mothers delivering premature infants was able to empirically show that administration of steroids saved infant lives and did not adversely harm the mother or child[71]. Prior to the publication of this SR, there was conflicting evidence about the efficacy and safety to mother and child of antenatal steroids. Today, administration of antenatal steroids is common practice, unquestioned, and has improved outcomes for premature infants. This SR was so influential that it has been immortalized as the logo for the

Cochrane Collaboration — an international group dedicated to conducting and publishing high-quality SRs on key clinical topics.

Just as with clinical trials, the benefit of SRs may be undercut by poor methods and design. One may argue that biased SRs are more harmful than biased clinical trials, since clinical trials are a form of primary research and may represent a single data point, whereas SRs are designed to resolve discrepancies in primary research and derive evidence-based summary effects. The means by which one ensures SRs are free from bias are largely similar to clinical trials: transparent registration, robust design and methods, and complete and availability reporting of all data. This section will address these topics and provide a primer on SR methods.

### *Registration*

Registering an SR offers similar benefits as what were discussed in the Trial Registration section above. That is, prospective SR registration allows one to not only claim their SR idea sooner than one could if they waited to publish their paper, but it also provides the public a time-stamped history of changes that can improve transparency in the SR process. Currently, the most commonly used SR registry is PROSPERO (International Prospective Register of SRs)[72]. Despite a growing number of SRs being registered in PROSPERO, it was recently shown that SR registration is much less common than clinical trial registration, with only 15.2% of SRs included in a recent study being registered[73]. The number of SRs whose registration was not up-to-date was staggering, with 85% not containing up-to-date information. The best use of an SR registry includes updating the information as an SR evolves. It may be the case that the proposed analysis is not possible, due to limitations in the primary research studies or unforeseen problems by the SR authors. It is likely that the lack of attention that SR registration has received compared to clinical trial registration is the source of the underuse of registries like PROSPERO. For ICMJE-member journals, trial registration is a condition for publication, but no such impetus exists for SRs[29]. Movements to improve the registration of SRs follow the same line of reasoning as for clinical trials: it is the best way to maximize trust and transparency in the SR process.

### *Design and methods*

One reason that SR registration is so important is because of the flexibility of the analyses. Statistically, SRs may be considered more complex than the average clinical trial, and there is a commensurate decrease in understanding of SR methods, which has prompted some researchers to publish step-by-step guides to understanding SR[74–76]. For those wishing to conduct an SR, it is recommended to adhere to the Cochrane Handbook of SRs[17] when determining study methods, and the PRISMA Statement (Preferred Reporting Items for SRs and Meta-Analyses)[77] when writing a manuscript. In basic terms, an SR consists of 4 sections: 1) literature search, 2) article screening, 3) data extraction, and 4) data analysis.

An SR literature search should be comprehensive and broad enough as to not miss any potentially relevant studies. It is recommended to search at least two different research databases, with the most common being PubMed (which includes MEDLINE) and EMBASE[77]. It is recommended to conduct the search with the help of a trained research librarian, since librarians are experts in database search syntax and optimization. In addition to searching at least two research databases, it is recommended to search "gray literature", also known as unpublished or yet-to-be published literature, and to augment one's search of the published literature by handsearching relevant medical journals or the citation lists of articles retrieved from the database search[78]. Searching gray literature mitigates the possibility of publication bias affecting the SR results. Publication bias occurs when only statistically significant results are published, which may exaggerate summary effects of interventions. Augmenting a search with handsearching methods mitigates the possibility that relevant articles are excluded because of suboptimal search terms or database indexing.

Article screening is relatively straightforward and involves at least two authors reviewing all articles retrieved from the literature search[79]. The best manner by which to screen articles is to keep all authors involved working separately without knowledge of the others' decisions. In doing so, bias may be prevented from affecting the inclusion or exclusion of studies. The authors judge each article according to prespecified inclusion criteria, which is commonly organized in the form of a PICO (population, intervention, comparator, outcome) question[80]. An example of a basic PICO question is randomized trials testing whether pembrolizumab [I] improves survival [O] compared to standard chemotherapy [C] in patients with lung cancer [P]. More advanced or narrow PICO questions are often constructed to improve the directness of an SR. A PICO question that is too broad may lead to unsatisfactory heterogeneity between studies. Heterogeneity is the degree to which patients, primary study methods, or primary study results vary from one other[81]. A heterogeneity statistic is often listed within a forest plot — the visual representation of meta-analytic results — and an assumption of whether heterogeneity is expected to exist is made prior to analysis. If heterogeneity is expected, the effects model for the meta-analysis may change. A balance between a PICO question that is broad and narrow enough to be clinically meaningful is the basis for a robust SR.

Data extraction follows a similar method to article screening and often includes at least two authors working separately without knowledge of the decisions of others[82]. What data is extracted may differ from SR to SR and is based on what information is necessary to answer the research questions that were posed. Data extraction must be comprehensive to allow for robust primary, subgroup, and sensitivity analyses, while refined enough so as to not distract from the key questions to be answered. A basic formula that would follow the PICO question posed above would be to extract the number of patients, the dose and administration schedule of pembrolizumab, the precise standard of care administered to control patients, the disease severity for included patients, whether other co-interventions were given (e.g., other medications), the efficacy outcomes (survival), and any safety data. In addition to data that is extracted to answer the research question, it is highly recommended that SR authors assess the included studies for risks of bias that may affect study results. For clinical trials, the use of the Cochrane Risk of Bias version 2 tool[83] is recommended, and for observational studies, the Risk Of Bias In Non-randomized Studies of Interventions tool is most appropriate[84]. The risk of bias of primary studies included in an SR may be the driver of a study's effect and is essential for understanding the results of an SR.

SR analyses come in three major forms: primary, subgroup, and sensitivity[74]. SRs are not powered for a primary outcome like clinical trials. In other words, the ability for an SR to truly detect an effect of an intervention is not contingent on the accrual of patient events or included studies. Rather, SRs define a primary outcome, which is akin to a primary question. In our PICO question above, OS would be the primary outcome and other outcomes, like safety analyses, would be secondary because they are more peripheral relative to our PICO question. Subgroup analyses divide included studies, or patients therein, into two or more groups to compare whether results differ. Variation in results may indicate that the overall effect is driven by the grouping variable that divides the cohort into its groups and the subgroup interaction analysis is used to statistically determine whether a difference exists between subgroups[85]. Sensitivity analyses are used to test the robustness of the data and are wide-ranging in their types[86]. The goal of a sensitivity analysis is to determine whether one or more studies was responsible for the SR summary effect. If the results of the SR are affected by a sensitivity analysis, then the overall results are considered fragile, unstable, and subject to question. Some common examples of sensitivity analyses include: removing one study at a time from the meta-analysis and re-analyzing results, removal of studies with high risk of bias, and removing studies based on study design (e.g., observational studies). Altogether, primary, subgroup, and sensitivity analyses are crucial to understanding the degree to which results are trustworthy, reliable, and applicable to the entire population of patients in question.

### Clinical Practice Guidelines

CPGs are consensus statements, developed by a group of multidisciplinary experts, which aim to guide healthcare practice and patient care in an evidence-based manner[87]. CPGs are often referenced by physicians seeking guidance on key aspects of patient care, insurance companies determining which interventions will be paid for, and patients seeking more information about their disease. Hundreds of CPGs exist, many of which overlap in their scope and topics, and are regularly updated as new evidence is published. CPGs are of the utmost importance to modern clinical practice, since they represent common ground for all stakeholders to gather and understand patient care. Because of the importance of CPGs, the Institute of Medicine (now the National Academy of Medicine) has published a lengthy document outlining best practices for CPG development and dissemination[88].

Several key methods must be implemented to ensure that CPGs are useful and robust, capable of identifying the most evidence-based treatments for common diseases.

To illustrate the extent of the problem in oncology that would have gone unnoticed if not for meta-research, one needs not look further than the National Comprehensive Cancer Network (NCCN) CPGs. The NCCN guidelines are the gold-standard for oncology practice in the United States, so much so that the NCCN guidelines are one of 5 compendiums for the Centers for Medicare and Medicaid (CMS) database[89]. In other words, if the NCCN recommends a drug, CMS is obligated to pay for it. The NCCN most often references clinical trials as its basis for its recommendations. This system works best for patients if the NCCN guidelines are evidence-based and free from potential bias. There is strong evidence from meta-research studies that NCCN authors hold large, highly-relevant conflicts of interest with drug manufacturers whose products are included in the NCCN guidelines[41]. There is additional evidence that rampant off-label use (i.e., non-FDA approved use) of drugs is recommended by NCCN authors[48]. That means that an oncologist can reasonably prescribe a drug to a patient without FDA approval and CMS, using tax dollars, is compelled to pay for it. Adding to all of this, the status quo of how oncologists are paid is by "cost-plus reimbursement" which pays oncology practices who purchase drugs from a manufacturer a percentage of the cost of the drug that is prescribed. This reimbursement procedure gives an incentive to prescribe more costly drugs[90]. Last, it has been empirically shown that when oncologists maintain conflicts of interest with a drug company, the oncologist is more likely to prescribe that company's drugs and more expensive drugs in general[91–93]. None of these facts would be understood without the use of meta-research methods.

### *Design and Methods*

The structure, methodology, and reporting of CPGs are best outlined by the Reporting Items for Practice Guidelines in Healthcare (RIGHT) statement[94] and the Appraisal of Guidelines, Research and Evaluation (AGREE-II) guidelines[95]. RIGHT and AGREE-II are considered the premier sources for those seeking to conduct or evaluate CPGs. These two documents outline the key CPG items that need to be reported and key methodological considerations that would ensure the CPG results and recommendations are robust. Even if these guidelines are not explicitly followed by CPG working groups, one may still find a high-degree of overlap between CPGs that are considered robust and the recommendations outlined by the RIGHT and AGREE-II documents.

The major reason why CPGs are considered to be the gold standard references for clinical practice is that they are based on an SR of the literature, evidence-based methods for rating the quality of evidence and applicability of interventions, and written by groups with access to experts in the field. The same principles apply to CPG authors that apply to SR authors: the best means to derive evidence-based recommendations is to use a systematic search of the published and unpublished literature to identify all relevant data. CPG development groups often have a team of librarians or search experts who conduct literature searches. These results are aggregated and interpreted by the clinical experts. These experts have a difficult task of interpreting literature that may be imprecise, indirect, or suffer from other risks of bias. Recall that many interventions in medicine are of small or modest effect sizes, which opens the possibility that favorable and unfavorable results for an intervention may coexist in the literature. Making sense of this tangled web of data is made easier using the GRADE (Grading of Recommendations, Assessment, Development and Evaluations) method of rating evidence[96].

The GRADE method is simple: one starts by assuming that all data is high-quality and is downgraded to moderate, low, or very low levels based on key factors. These key factors cover a multitude of nuanced methodological shortcomings, but broadly, the GRADE method downgrades evidence for 1) imprecision, 2) indirectness, 3) inconsistency, and 4) risk of bias[96]. Imprecision refers to data that shows a wide range of variability, often due to insufficient sample size to obtain a robust result[97]. Indirectness refers to data that is somewhat relevant to the PICO question posed, but certain aspects are tangential[98]. An example of data that is indirect is a trial of a relevant intervention against the relevant comparator, but in a population that differs from the one of interest. Reasonable conclusions may be drawn about the intervention's effectiveness based on clinical gestalt, but nonetheless the answer is unknown. Inconsistency refers to data sets that compete with one another, as in the case of two trials where one shows a significant effect in favor of the intervention and one shows no such effect[99]. Risk of bias is best identified using robust tools[83,84,100] mentioned in the SR section above, and may reflect a range of methodological flaws that can undermine trustworthy results[101]. Levels of evidence are often shown as a numerical grade – 1 for high-quality with increasing numbers representing a step down in methodological quality.

These alphabetical scores for levels of evidence are often combined with numerical scores for clinical recommendation grade, as is done in the European Society of Medical Oncology (ESMO) guidelines[102]. The clinical recommendation grade is essentially the CPG authors declaring how assured they are that a therapy or practice should be incorporated into clinical practice. This combination of alphanumeric scores is a shorthand for whether evidence is robust and whether the data indicates the practice should be implemented. A variety of combinations may be used, such as 1A, which indicates high-quality evidence that an intervention should be used, and 1D which indicates high-quality evidence that a practice should be avoided. There is significantly more flexibility in how CPG authors assign clinical recommendations than in how they assign levels of evidence GRADE scores. CPG authors may use a combination of level of evidence, clinical gestalt, and patient values and preferences to justify a clinical recommendation. It goes without saying that the best means by which to ensure the process of deriving CPG recommendations is to form a multidisciplinary team of experts who are free from undue external influences and fully equipped to make the right recommendations. Cases where external influences on CPGs have occurred have been disastrous and likely resulted in patient harm[103], especially because some prominent CPGs are used to guide insurance coverage and reimbursement.

### ***Bias and Reproducibility***

Now that a clear understanding of the background and importance of clinical trials, SRs, and CPGs is understood, it is important to know what forms of bias may affect these three forms of research output. The rigor and reproducibility of oncology evidence is directly affected by suboptimal methods, interpretation, reporting, and dissemination of research. This section will aim to define common terms, explore consequences of various forms of bias, and demonstrate real-world consequences of these biases. A general overview will be followed by a more narrow, tailored exploration of bias in oncology research.

## Clinical Trials

### *Design*

The most appropriate manner by which one may structure a discussion on bias in the design of clinical trials is via the PICO framework discussed previously. Therefore, the first area of interest is the Population that is enrolled in a clinical trial. Clinical trials are also known as “controlled trials” because of their highly rigid and structured set of rules and inclusion criteria. These rules and criteria are meant to reduce any noise or error in the measurements of the efficacy of a cancer therapy, but have led to trial populations being different in terms of age, gender, type and location of cancer, severity of disease, and postoperative treatments[104–107]. The criteria are also meant to increase the rate of trial enrollment, since younger and healthier patients are more likely to comply with cancer trial protocols that require regular follow up[108]. On the contrary, it is no surprise that the strict rules leading to younger and healthier patients in cancer clinical trials has led many to question how the trial efficacy translates to real-world effectiveness. This lack of translation was evident with sorafenib for the treatment of hepatocellular carcinoma after primary resection. In the trial, sorafenib had a 2.8-month survival advantage over placebo, but the patients had more vitality and earlier-stage liver disease than real world patients. In the follow up, propensity-score matched analysis of patients in the real-world, the effect of sorafenib was erased, likely because the patients who received sorafenib in the real world were older and sicker than the trial participant.

With respect to cancer trial interventions, there are few questions about the dosage or mode of administration. A more fundamental question was recently raised about why certain drugs are being studied in the first place. The natural progression of a novel cancer therapy is to first test it in a Phase 1 trial to establish the safety and most appropriate dose, then move to Phase 2 where efficacy of a drug is established based on a drug’s ability to halt cancer growth or eradicate cancerous cells. Only after Phase 1 and 2 trials are favorable drugs moved to Phase 3, where survival and quality of life are tested. A recent search of *New England Journal of Medicine*, *Journal of Clinical Oncology*, *Lancet*, and *Lancet: Oncology* — the four most prominent publishers of cancer clinical trials — found that "negative" Phase 3 trials of novel cancer therapies published in 2016 were not supported by Phase 2 evidence half of the time[109]. In other words, companies proceeded to Phase 3 even if the Phase 2 data was negative or inconclusive. The result was a nonsignificant and unfavorable result in Phase 3. The choice to proceed to Phase 3 is postulated to be based on either the "sunk cost bias", in which companies have pursued a drug further to try and compensate for potential losses, or the gamesmanship of Phase 3 trials[109]. In particular, this gamesmanship takes the form of the fact that Phase 3 trials enroll more patients, and are therefore more statistically powerful and able to detect smaller differences than a Phase 2 trial. So, companies may pursue Phase 3 testing knowing that a smaller difference will be identified and can be used to lobby for FDA approval. The overall result may be that drugs with clinically insignificant benefits are approved because of statistical gamesmanship, not the drug's true value.

Oncology clinical trial control arms are a more contentious area of research. To understand the scope of the problem, it is important to understand how oncology trials function globally. The United States drug market is the most lucrative in the world[110], and it is therefore the most sought after by for-profit companies. Demonstrating efficacy across the globe is important for drug approval and use in other countries. It is common for modern oncology trials to enroll patients globally[111]. Given the cost of these novel cancer drugs, it is unlikely that anyone that lives in a low- or middle-income country will be able to afford the trial drug after it is on the open market. Access to that drug for trial participants is clearly beneficial if the drug is effective. The caveat is that these low- and middle-income countries also cannot afford recently approved cancer drugs that may be standard of care at the time of the novel trial. That means that unless the for-profit company buys and donates the standard of care drug that is used as a control arm in low- and middle-income trial centers, those trial centers are instructed to use the best available control arm drug at their disposal. This was empirically studied recently and it was found that 17% of new FDA-approved drugs between 2013 and 2018 were based on a suboptimal control arm[39].

The implication is that approximately 1 in 5 new drugs approved in the United States, the perceived efficacy may be lower than what is true.

Finally, oncology clinical trial outcomes are the most robust source of confusion and bias. There are, broadly, two types of endpoints in oncology and medical research: clinical endpoints and surrogate endpoints. Clinical endpoints directly measure patient outcomes, and some examples include survival, quality of life, or major adverse cardiac events (in the case of cardiovascular trials). Surrogate endpoints indirectly measure clinical endpoints and examples include delay in tumor growth, decreased tumor burden, or changes in a lab value (e.g., cholesterol). Many drugs have anticancer activity, and may delay tumor growth or decrease tumor size, but do not improve survival. This is difficult to understand because many consider shrinking a tumor as a fundamentally helpful thing for patients. Indeed, it is when it can provide symptomatic relief to patients. However, many patients do not present with symptoms of their tumor being too large and causing pain. For example, the most common presenting symptom for lung cancer is cough[112], and lung cancer often presents in advanced stages, sometimes with metastatic disease[113]. Surrogate endpoints that measure delay in tumor growth set arbitrary growth criteria that have no basis in patient symptoms, pain, or discomfort[114]. A delay in tumor growth may not be truly valuable to patients unless it corresponds with improvements in survival. As it turns out, repeated analyses show that such surrogate endpoints do not correlate with survival or quality of life, as would be expected if symptomatic relief were conferred[61,62,115]. Nonetheless, surrogate endpoints are heavily valued in oncology clinical trials and are most often the basis for FDA approval of novel therapies[3].

### *Analysis and Interpretation*

Sample size calculations are fundamental to the analysis and interpretation of a clinical trial. A sample size calculation consists of 3 items: the estimated effect size, the type 1 error rate (i.e., alpha), and the type 2 error rate (i.e., beta). The estimated effect size should ideally be based on previous literature and may come from an observational study, phase 2 trial, or any other robust piece of literature that supports the estimated effect that one would expect to see for the intervention. The alpha percentage is often defaulted at 0.5, or 5%, to represent a one in twenty chance that the observed effect in the trial would be that large if the intervention were truly ineffective. In other words, imagine a trial in which a novel cancer therapy reduced the absolute risk of death by 20%, $P = .05$. Since we start with the null hypothesis assumption that the intervention has no effect, it would not be expected to show an effect that large more than 5% of the time if it were truly ineffective. Therefore, all else being equal, the fact that it demonstrated a large effect indicates that it is a truly effective drug. The type 2 error rate, or beta, is colloquially referred to as the "false negative rate", or the rate at which we would fail to identify an effect that truly exists. Beta is important because 1 - beta gives us the study power, which is the probability that the study has of identifying a true effect when it exists. The effect size, alpha, and beta are used to generate a sample size that is necessary to conduct the trial under the statistical assumptions laid out.

Modern oncology trials funded by for-profit companies have the resources to recruit more patients than trials of previous decades. It is well known in scientific research that a larger sample size is required to identify statistically significant small effect sizes. A recent study by Ocana and Tannock reviewed trials approved by the FDA for novel cancer therapies and found that in some the observed effect was smaller than the estimated effect in the sample size calculation. The implication is that the observed effect is smaller than what the authors considered clinically significant enough to include in a sample size calculation, yet because the observed effect was statistically significant, the trial was "positive". Another paper reviewed the mean effect size of novel cancer therapies and found that the average cancer drug improves survival by 2.4 months[116].

Hazard ratios are the default effect size used in the majority of oncology clinical trials because hazard ratios are a measure of instantaneous risk to an individual, rather than an estimate of risk to a group as is seen with odds ratios. Hazard ratios neatly fit into the structure of oncology trials, which rely heavily on survival analyses and Kaplan-Meier curves. A hazard ratio is difficult to interpret and translate to clinical benefit, since it does not capture the overall, longitudinal risk to a patient[117–119]. Additionally, hazard ratios are calculated using models that are difficult to understand and rely on a set of assumptions that, when violated, render hazard ratio estimates untrustworthy. The key assumption worth discussing further is the assumption that censoring of a trial participant is unrelated to the prognosis of that individual in the trial.

Censoring has recently become a hot topic in oncology research. Censoring in a clinical trial happens when an individual experiences the outcome event or they are lost to follow up[120]. A censored participant is represented by tick marks on a survival curve. Imbalances in censoring may occur for multiple reasons, but the most dangerous reason is due to physician knowledge of the control arm patients. Some side effects in oncology are notably more common in some drugs or drug class than others[121]. If an oncologist notices a side effect associated with the control arm, they may heighten their clinical investigation of the patient leading to earlier detection of an event of interest (e.g., tumor growth) in the control arm. As surrogate endpoints are so popular and common, the majority of new trials are subjected to this bias. An empirical analysis of differential censoring found that censoring was more common in the control arm early in trials, which is indicative of heightened clinical investigation[122].

Another driver of heightened clinical investigation early in clinical trials is the fact that crossover is built into trial designs. Crossover occurs when a patient in one arm crosses over to receive the therapy from the other arm[49]. In oncology trials, this crossover occurs unidirectionally from the control arm to the intervention. Not all oncology trials are completely blinded, either because one arm has oral medications and the other has intravenous, but also because of the aforementioned common side effects in a certain class of drugs. A physician who knows the trial is comparing a novel immunotherapy to an older chemotherapy, and may not only be able to tell which therapy a patient has received, but also favor the novel intervention and want a patient to receive that therapy. In such cases, crossover may be a sought-after phenomenon by a physician. Crossover can taint the analysis of clinical trials because a patient is often analyzed in the group they were assigned, regardless of whether they crossed over[123,124]. So, if patients in the control arm are moved to the intervention, it may prolong their life and obscure the true effect of an intervention and result in a false-negative or diminished effect size. This crossover may be okay if a drug has already established its baseline efficacy in a previous trial, but many novel drugs, without baseline proof of efficacy allow crossover. In such cases, regulators, oncologists, and patients may not know the true effect size going forward. The proper and improper use of crossover has been thoroughly discussed by experts in the field to provide a framework for future trial to follow[125].

Last, the use of interim analyses in oncology trials may subject trials to biased interpretations and analysis. An interim analysis is essentially an "early look" at the trial data, and if the data is favorable enough, the trial may be stopped early for benefit. Similarly, if the data is unfavorable enough the trial may be stopped early for harm[126]. Strict statistical criteria are used to keep trials free from bias in traditional interim analyses[127]. In oncology trials interim analyses have transformed into an analysis of surrogate endpoint data alone. Most oncology clinical trials have 2 primary endpoints, each with their own sample size estimation, and enrollment occurs continuously until both endpoints have accrued enough events. One of these primary endpoints is a surrogate endpoint, for which events accrue more quickly, and the other is often OS, for which events accrue slowly. An oncology trial will have fully matured data for the surrogate endpoint months before OS, and it is common for trials of novel oncology drugs to publish the results of their interim analysis of surrogate endpoint data months before the OS data has fully accrued events. This was studied outside of oncology and it was found that the conclusions of interim analyses changed 21% of the time[128]. In oncology, there are open questions about whether interim results with surrogate endpoints will translate to OS benefit, and whether the subsequent publications receive as much hype and attention in the oncology community.

### *Reporting*

Several key reporting biases have been well documented across medical research. The first one worth understanding is publication bias. Publication bias occurs when trials that obtain unfavorable, often not statistically significant results, are published at a lower rate than trials that obtain favorable, often statistically significant results.[129] There are two main drivers of publication bias — author-driven and journal driven. Author-driven publication bias occurs when authors are unsatisfied or uninterested in the results of their study and decide to not pursue publication. Journal-driven publication bias occurs when journals reject papers that do not include statistically significant results, perhaps because these results are not attractive to readers. The effect of publication bias is a portfolio of published research that is overwhelmingly favorable to the intervention. The seminal paper on publication bias compared the portfolio of results for antidepressants in the published literature and FDA drug efficacy reviews, where submission of all clinical data is compulsory.[129] Authors of this paper found 74 studies submitted to the FDA, of which 31%, including 3,449 patients' data, were not published in medical journals. In the published literature, 94% of antidepressant trials were "positive", whereas only 51% of studies submitted to the FDA were "positive". The implication is a biased portfolio of research available to everyday physicians and readers, and downstream effects that may occur when these studies are aggregated in meta-analyses or CPGs.

Selective outcome reporting bias is another form of bias that affects the portfolio of medical evidence. Selective outcome reporting occurs when authors of medical research papers include, exclude, or change study endpoints on the basis of statistical significance[130]. There have been massive gains in the understanding and monitoring of selective outcome reporting bias, thanks in large part to the advent of clinical trial registries[34], federal policy that requires prospective trial registration[131], and the open data movement[132]. Without these advancements, medical research would be subjected to investigators' or study sponsors' desires to publish a positive and wholly favorable study. The downstream effects of selective outcome reporting bias include changes to patient care based on the perception of highly favorable results, exaggeration of meta-analysis effect sizes, and distrust of scientific research. The prevalence of selective outcome reporting bias has been empirically studied across medicine[133–135], including hematology[36] and oncology[136].

Spin is a subtype of selective outcome reporting bias, and is subtler and more subjective. Spin is defined as "use of specific reporting strategies, from whatever motive, to 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"[137]. The impact of spin goes beyond medical research papers. Evidence of spin has been found in press releases[138] and even peer review comments sent to the author[139]. The effect of spin on physicians was evaluated in a group of 300 oncologists, of whom half were randomized to the spin group and the other was randomized to the no spin group[52]. In the spin group, 30 abstracts with at least one type of spin were given to oncologists to read. The types of spin include selective outcome reporting, omission of unfavorable results from the discussion, and emphasis on a subgroup that performed well. In the no spin group, these 30 abstracts were rewritten by the study authors to have no spin or distortions, then given to each oncologist to read. The group that read the abstracts with spin were more likely to rate the intervention as more beneficial, say they would read the whole article, and rate the trial as less rigorous. The implication is clear: misrepresented and distorted research findings in medical research affect how physicians perceive the efficacy of drugs that are studied.

The common thread for how to prevent all reporting biases is to emphasize open data, pre-registered protocols, and standardized reporting guidelines. If all trials were registered with their planned endpoints, evidence of publication and selective outcome reporting bias is easy to identify. If standardization to how studies are reported is emphasized, spin would become less common. In all cases, studies would be independently reproducible if data and statistical analysis plans were made available at the time of publication. The Open Science movement[132] is predicated on equity for patients, the public, and medical researchers in terms of access to data that impacts how clinical decision making is determined.

## Systematic Reviews

### *Design*

The success of an SR begins with a robust and comprehensive database search based on a well-defined PICO question. Bias in an SR occurs when the search neither maximizes its sensitivity and specificity[17]. SRs that do not search multiple, relevant databases are unlikely to identify all relevant studies. SRs that do not search for unpublished data may not generate reliable results[78]. Unpublished data, or gray literature, is more likely to have smaller or null effect sizes because of publication bias. Inclusion of only the published literature may result in exaggerated effect sizes in a direction that is favorable to the intervention. SRs that do not include all languages may or may not generate unreliable results. There is a debate about whether or not including only English-language studies is appropriate, with advocates of including English-language articles highlighting the fact that non-English studies are more likely to be at a higher risk of bias due to financial restrictions of low- and middle-income countries[140]. This does not seem to be an axiom and it is likely that the effect of language bias is sporadic, though still present. To that effect, the Cochrane collaboration recommends minimizing language bias as best as possible within the bounds of budget and time restraints[17].

A particularly common and unrecognized form of bias in SR searches centers on what is known as citation bias. Citation bias occurs when authors of SRs try to bolster their search returns by hand-searching the reference lists of relevant articles identified in the SR search[141]. This may seem logical until one understands that the manner in which studies are cited is often biased. Cited studies are more likely to confirm the results of the study doing the citing, and cited studies are subjected to publication bias, meaning they are more likely to have statistically significant results. Therefore, the practice of identifying new studies from the reference lists of included studies may be seen as a method to reduce bias, but may in fact be increasing the bias in an SR. The prevalence of citation bias in the SRs in the field of otolaryngology was recently studied and it was found that
72.4% (390/539) hand-searched reference lists of included articles with 58.5% (228/390) of those SRs not including any other form of gray literature search.

*Reporting*

In a similar fashion to clinical trials, SRs are subject to publication bias, selective outcome reporting bias, and spin. The manner by which each bias is mitigated is similar to clinical trials as well: pre-registration, publicly available protocols with statistical analysis plans, and guidelines for how data should be reported and interpreted. Unfortunately, the rate of pre-registration of SRs is far lower than that of clinical trials[73]. The availability of protocols is low, with some estimates finding that just under half of SRs have public protocols[142], while others finding closer to one-third[143]. One can imagine the effect that lack of pre-registration and public protocols would have on the published SR literature, and indeed, a recent meta-analysis of studies investigating selective outcome reporting bias found that 38% (184/485) included, excluded, or omitted outcomes[82]. It is likely that this number is a lower end estimate because it is only possible to study selective outcome reporting if a protocol or preregistration is available. Publication bias has been quantified, with past studies finding that the SRs in general medical journals differ from those in the Cochrane collaboration journal, which requires publication regardless of results[144,145]. Spin, otherwise known as the distortion or misrepresentation of research findings, was recently studied in the abstracts SRs of breast cancer[146], orthopedic[147], and glaucoma[148] interventions — each study finding a significant amount of spin that may affect the interpretation of SR results.

### Clinical Practice Guidelines

At a fundamental level, CPGs are an SR. All relevant literature is retrieved for a certain topic, but unlike normal SRs, CPGs rely on an expert panel to issue clinical practice recommendations based on the available evidence. CPGs may or may not synthesize results from included articles, but more often issue qualitative recommendations based on the preponderance of evidence gathered[88]. It is imperative that CPGs are based on a robust, systematic database search, that authors are experts in evidence synthesis, and that experts are free from any bias that could affect their recommendations. Beyond those basic tenets of CPG development, other, more nuanced, methodological items should be considered. These additional items are best summarized by the AGREE-II (Appraisal of Guidelines for Research and Evaluation version 2) tool, which is commonly used to assess the methodological quality of CPGs. The AGREE-II instrument asks whether a CPG 1) defines its scope and purpose; 2) outlines stakeholder involvement; 3) has high methodological quality; 4) clearly presents results and recommendations; 5) is applicable to clinical practice; 6) is free from external conflicts of interest. In oncology, assessments using AGREE-II have consistently shown that oncology CPGs do not sufficiently involve key stakeholders, including patients, nor are they free from conflicts of interest[149–151].

Without patient and other non-physician stakeholder involvement, it is unlikely that CPGs will consistently make recommendations that align with public interest. For example, most novel oncology drugs are so costly that a new term has been coined in oncology: financial toxicity[6]. Financial toxicity is unlike other drug toxicities, which occur when someone takes a drug. Financial toxicity precludes someone from taking a drug because the cost is so burdensome that the individual cannot afford it. Thus, it may be that patients and other relevant stakeholders would prefer an older drug that is more cost-effective, because a full course of an older drug may be more effective than half a course of a newer drug. Some have questioned whether financial conflicts of interest play a role in the clinical recommendations of oncology CPG authors[41]. Evidence suggests that oncology CPG authors have extensive financial COI with companies who develop novel cancer drugs[41]. Some would suggest that these conflicts are to be desired since a close working relationship with industry is working to serve the patient's best interest[152]. However, ample evidence suggests that financial conflicts of interest not only affect guideline recommendations[153], but also physician prescribing behavior[154] and interpretation of clinical data[155].

### *Summary*

Clinical trials, SRs, and CPGs are vital to medical decision making in oncology. A strong understanding of clinical trials, SRs, and CPGs is necessary to appreciate the findings of this dissertation. Furthermore, a baseline understanding of bias is important since many of the studies included in this dissertation discuss the prevalence, influence, and implications of bias from the perspective of research translation to clinical practice. Without improvements in the rigor and reproducibility of oncology evidence, which include improvements in study design, stakeholder involvement, and bias mitigation, it is unlikely that future oncology evidence will maximize its potential to improve patient outcomes.

# CHAPTER III

# ADHERENCE TO REPORTING GUIDELINES AND CLINICAL TRIAL REGISTRATION POLICIES IN ONCOLOGY JOURNALS: A CROSS SECTIONAL REVIEW

***This work was previously published in the* British Medical Journal: Evidence Based Medicine *with the following citation:***

Wayant C, Moore G, Hoelscher M, Cook C, Vassar M. Adherence to reporting guidelines and clinical trial registration policies in oncology journals: a cross-sectional review. *BMJ Evid Based Med*. Published online April 13, 2018. doi:10.1136/bmjebm-2017-110855

---

## *Introduction*

Poor methodological quality and incomplete reporting of published research affects clinical decision making [156–158] and contributes to research waste [159]. Reporting guidelines (RGs) offer one solution by promoting transparency and ensuring that key methodological safeguards are fully reported [160]. In oncology, recent studies have shown evidence of poor reporting quality of phase II and III trials [161,162]. Moreover, a survey of members from the European Organization for Research and Treatment of Cancer found that the frequency of adverse event reporting fell short of members' expectations [162]. To address such situations, correct implementation of the Consolidated Standards of Reporting Trials (CONSORT) checklist by clinical trialists would ensure that all harms and unexpected effects encountered by the treatment group are reported in oncology trials (CONSORT Item 19) [163]. Indeed, the CONSORT statement, like other RGs has been shown to improve the quality of research when incorporated into study design and reporting.[164,165]

The registration of clinical trials is another mechanism intended to promote transparency and improve methodological standards. Two recent studies demonstrated small improvements over time in the quality of trial registration, but conclude that more improvement is necessary with respect to important items such as clearly defined primary outcomes.[166,167] Discrepancies between outcomes listed in the trial registry record and those reported in the published trial have been noted across many medical specialties, providing indirect evidence for selective reporting bias. This bias occurs when researchers preferentially include (or exclude) outcomes based on statistical significance (or a lack thereof) [82,168]. The International Committee of Medical Journal Editors (ICMJE) and the World Health Organization (WHO) have instituted policies to improve clinical trial reporting and registration. The US government has made prospective clinical trial registration a legal mandate [169], and similar regulations have been implemented in Europe [170]. In January 2017, the US National Institutes of Health (NIH) began requiring registration of all NIH-funded randomized trials in ClinicalTrials.gov (the US clinical trial registry) prior to patient enrollment and reporting of summary results after trial completion [171].

In this study, we first evaluated the published guidance (e.g., instructions for authors) provided by a cohort of highly ranked oncology journals to authors regarding the use of RGs for common study types. We also examined these journals' policies on clinical trial registration. We then evaluated whether this guidance has led to improvements in reporting and registration.

### *Methods*

The primary outcome of this study was to examine the adherence to RGs and trial registration policies of 21 oncology journals. The secondary outcome was to investigate whether adherence to the CONSORT statement and ICMJE trial registration policies affects reporting practices in oncology. Our exploratory outcome was a description of the rates of adherence to oncology-specific RGs (e.g., REMARK for tumor marker prognostic studies).

We surveyed Google Scholar and identified the top 20 oncology journals, sorted by h5-index. We also included *JAMA Oncology* because its impact factor places it in the top

20 oncology journals, but Google rankings do not yet reflect that status. We conducted a cross-sectional review of the oncology journals' policies and instructions for authors concerning guideline adherence and trial registration requirements. This study did not meet the regulatory definition of human subject research, so it was not subject to Institutional Review Board oversight. We applied relevant SAMPL guidelines for reporting descriptive statistics [172].

Before initiating the study, all authors met to outline the study design. A study protocol was developed based on our previous investigations [173,174]. A pilot test was done on the first 5 journals to identify any flaws in the protocol and to establish uniformity in data extraction. Follow-up meetings were held periodically throughout the data extraction process to resolve discrepancies. CW, MH, and CC performed web-based searches for each journal and searched the instructions for authors page for relevant information. A third author (GM) validated all data. Each author was blinded to the ratings of the others. The methodology for the following Primary and Secondary outcomes is visual depicted in Figure 1.

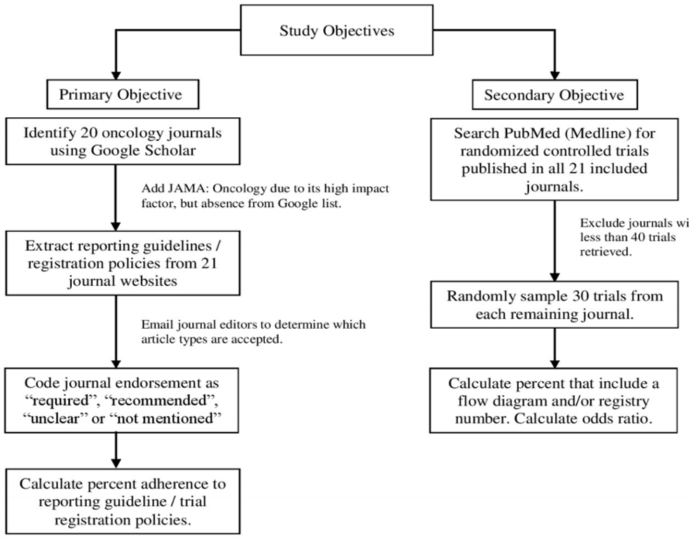
Study Objectives
Primary Objective
Secondary Objective
Identify 20 oncology journals using Google Scholar
Add JAMA: Oncology due to its high impact factor, but absence from Google list.
Extract reporting guidelines / registration policies from 21 journal websites
Email journal editors to determine which article types are accepted.
Code journal endorsement as "required", "recommended", "unclear" or "not mentioned"
Calculate percent adherence to reporting guideline / trial registration policies.
Search PubMed (Medline) for randomized controlled trials published in all 21 included journals.
Exclude journals with less than 40 trials retrieved.
Randomly sample 30 trials from each remaining journal.
Calculate percent that include a flow diagram and/or registry number. Calculate odds ratio.

*Figure 1. Study methodology for the primary and secondary objectives.*

*Primary Outcome*

For each journal, we determined whether it adhered to ICMJE Uniform Requirements for Manuscripts (URM), Animal Research: Reporting of In Vivo Experiments (ARRIVE), Case Reports (CARE), CONSORT, Meta-Analysis of Observational Studies in Epidemiology (MOOSE), Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA), Quality of Reporting of Meta-analyses (QUOROM), Strengthening the Reporting of Observational Studies in Epidemiology (STROBE), or Standards of Reporting Diagnostic Accuracy Studies

(STARD). Additionally, we extracted whether or not a journal mentioned ClinicalTrials.gov, WHO trial registries, or both. If a journal mentioned trial registration without naming a specific registry, we coded that journal as “generic trial registration.”

Definitions were constructed *a priori* by CW and MV for the coding process based on previous literature definitions [173,174]. For each of the RGs and registries, adherence by each journal was classified as “compulsory/required,” “recommended,” or “not mentioned.” For cases in which it was unclear whether the journal followed a specific guideline or registry, adherence was rated as “unclear.” Keywords such as “must,” “need,” or “manuscripts will not be considered for publication unless” were categorized as compulsory/required. Similarly, keywords such as “should,” “encouraged,” and “prefer” were categorized as recommended.

After data extraction, MH and CC reviewed each journal’s website to determine which of the common study types relating to extracted RGs (systematic reviews/meta-analyses, clinical trials, diagnostic accuracy studies, case reports, epidemiological studies, and animal studies) were accepted. Next, MH and CC emailed the editors-in-chief of the included journals for confirmation regarding the extracted study type. We sent two reminder emails at 1-week intervals to ensure best practices in eliciting email response [175]. We cross-referenced the journals’ accepted article types to the data that we extracted from journal websites. If a journal did not publish a particular type of study, then it was not considered in comparing accepted study types and RG adherence. For example, ARRIVE guidelines were not considered relevant to a journal if it did not publish preclinical animal studies.

### *Secondary Outcome*

Next, CW performed a PubMed search using publication type “randomized controlled trial” for all included journals during a 5-year period (January 1, 2012, to December 31, 2016). This search strategy has been shown to have over 93% sensitivity and specificity for retrieving RCTs [176]. All RCTs were divided into groups based on whether or not the journal adhered to CONSORT guidelines and whether or not they endorsed ICMJE trial registration policies. CW then randomly sampled 30 RCTs from each journal. If a journal did not publish at least 40 RCTs during the 5-year study period, it was excluded. Odds ratios and confidence intervals were calculated based on the results of the data extraction using STATA 13.1.

*Exploratory Outcome*

A single author (CW) surveyed each journal's website to descriptively analyze the rate of adherence to oncology-specific RGs. A list of these guidelines can be found on the EQUATOR Network's website [177].

All authors met after completing data extraction and analysis to resolve any final discrepancies in the scoring of the journal data.

***Results***

Our study comprised 21 oncology journals. Table 1 shows all extracted data. Only 1 (4.8%) journal was found to not adhere to any RG, while 5 (23.8%) did not adhere to any trial registration policies. The ICMJE-URM was mentioned by 15 (71.4%) journals, and the EQUATOR Network was mentioned by 3 (14.3%) journals. We recorded an editor response rate of 52.4% (11 of 21).

Primary Outcome

*Reporting Guideline Adherence*

The CONSORT statement was mentioned by 16 journals: 11/21 (52.4%) required adherence, and 5/21 (23.8%) recommended adherence. ARRIVE was mentioned by 11 journals: 1/20 (5.0%) required adherence, and 11/20 (55.0%) recommended adherence. STARD was mentioned by 8 journals: 4/19 (21.1%) required adherence, and 4/19 (21.1%) recommended adherence. PRISMA was mentioned by 8 journals: 3/21 (14.3%) required adherence, and 5/21 (23.8%) recommended adherence. STROBE was mentioned by 7 journals: 5/21 (23.8%) required adherence, and 2/21 (9.5%) recommended adherence. MOOSE was mentioned twice and was recommended both times. QUORUM and CARE were not mentioned in any journal's instructions for authors page.

*Trial Registration*

Five (23.8%) journals did not mention trial registration at all. Ten journals mentioned trial registration through ClinicalTrials.gov: 3/21 (14.3%) required registration, and 7/21 (33.3%) recommended it. Six journals mentioned WHO trial registration: 4/21 (19.0%) required registration, and 2/21 (9.5%) recommended it. Generic trial registration was mentioned by 11 journals: 8/21 (38.1%) required generic registration and 3/21 (14.3%) recommended it.

## Secondary Outcome

Our PubMed search yielded 2614 results and 13 eligible journals, defined as those publishing more than 40 RCTs in the 5-year period studied. Journal specific publication rates of a CONSORT Flow Diagram and trial registry number are available via the Open Science Framework (osf.io/7g6td).

*CONSORT Guidelines*

Of the 13 journals, 10 adhere to CONSORT guidelines and 3 do not (Figure 2). The RCTs published in the 10 journals that adhere to CONSORT included a flow diagram 70.3% (211/300) of the time. The 3 that do not adhere to CONSORT included a flow diagram 57.8% (52/90) of the time. This finding indicates that journal adherence to CONSORT increases the likelihood of an author adhering to its key items (OR=1.73, 95% CI: 1.03-2.89).

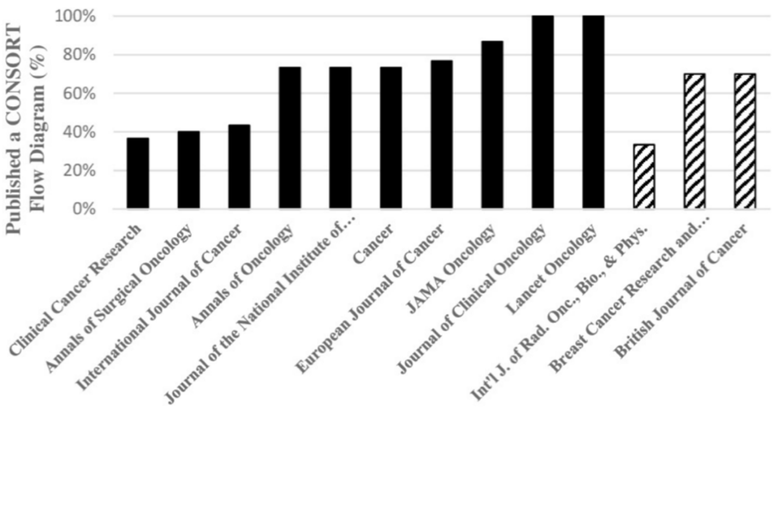

*Figure 2. Histogram of percent of randomly sampled randomized controlled trials (RCT) that included a Consolidated Standards of Reporting Trials (CONSORT) flow diagram published in journals that adhere to the CONSORT guidelines (solid) and those that do not (striped).*

| Journal Title | EQUATOR | ICMJE | CONSORT | MOOSE | QUOROM | PRISMA | STARD | STROBE | ARRIVE | CARE | REMARK | CT.gov | WHO | Other Registration |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Journal of Clinical Oncology | No | Yes | 2 | 4 | 4 | 4 | 4 | 4 | 4 | N/A | Yes | 2 | 4 | 1 |
| The Lancet Oncology | Yes | Yes | 1 | 4 | 4 | 2 | 1 | 1 | 4 | N/A | No | 1 | 1 | 4 |
| Cancer Research | No | Yes | 1 | 4 | 4 | 1 | 1 | 1 | 2 | 4 | No | 2 | 1 | 4 |
| Nature Reviews Cancer | No | Yes | 1 | 4 | 4 | 4 | 4 | 4 | 4 | N/A | Yes | 1 | 4 | 1 |
| Cancer Cell | No | No | 4 | 4 | 4 | 4 | 4 | 4 | 2 | 4 | No | 4 | 4 | 4 |
| Clinical Cancer Research | No | Yes | 1 | 4 | 4 | 1 | 1 | 1 | 2 | 4 | No | 2 | 1 | 4 |
| Annals of Oncology | No | No | 1 | 4 | 4 | 4 | N/A | 4 | 4 | N/A | No | 4 | 4 | 4 |
| Oncogene | No | Yes | 1 | 4 | 4 | 4 | 4 | 4 | 2 | N/A | No | 2 | 4 | 1 |
| Cancer Discovery | No | Yes | 1 | 4 | 4 | 1 | 1 | 1 | 2 | 4 | No | 1 | 2 | 4 |
| Cancer | No | Yes | 2 | 4 | 4 | 4 | 2 | 1 | 4 | N/A | No | 2 | 1 | 4 |
| Journal of the National Cancer Institute | No | Yes | 1 | 2 | 4 | 2 | 2 | 2 | 2 | 4 | Yes | 4 | 4 | 2 |
| Leukemia | No | Yes | 1 | 4 | 4 | 4 | 4 | 4 | 1 | N/A | No | 2 | 4 | 1 |
| British Journal of Cancer | No | No | 4 | 4 | 4 | 2 | 4 | 4 | 4 | N/A | No | 4 | 4 | 2 |
| International Journal of Cancer | Yes | No | 2 | 4 | 4 | 2 | 2 | 2 | 2 | N/A | No | 4 | 4 | 2 |
| International Journal of Radiation Oncology * Biology * Physics | No | Yes | 4 | 4 | 4 | 4 | 4 | 4 | 2 | 4 | Yes | 4 | 4 | 1 |
| Cancer Letters | No | No | 4 | 4 | 4 | 4 | 4 | 4 | 2 | 4 | No | 4 | 4 | 4 |
| European Journal of Cancer | No | Yes | 2 | 4 | 4 | 4 | N/A | 4 | 2 | 4 | Yes | 4 | 4 | 1 |
| Annals of Surgical Oncology | No | Yes | 2 | 4 | 4 | 4 | 4 | 4 | 4 | N/A | No | 4 | 4 | 4 |
| Nature Reviews Clinical Oncology | No | Yes | 1 | 4 | 4 | 4 | 4 | 4 | 4 | 4 | No | 1 | 4 | 1 |
| Breast Cancer Research and Treatment | No | No | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 4 | Yes | 4 | 4 | 4 |
| JAMA Oncology | Yes | Yes | 1 | 2 | 4 | 2 | 2 | 4 | 4 | N/A | No | 2 | 2 | 1 |

*Table 1. Cross tabulations of oncology journals and the adherence to reporting guidelines and trial registration policies*

Key: 1 (required), 2 (recommended), 3 (unclear), 4 (not mentioned), N/A (does not accept the corresponding study type).
ARRIVE, Animal Research: Reporting of In Vivo Experiments; CARE, Case Report; CONSORT, Consolidated Standards of Reporting Trials; ICMJE, International Committee of Medical Journal Editors; MOOSE, Meta-Analysis of Observational Studies in Epidemiology; NA, not applicable; PRISMA, Preferred Reporting Items for Systematic Reviews and Meta-Analyses; QUOROM, Quality of Reporting of Meta-analyses; STARD, Standards of Reporting Diagnostic Accuracy Studies; STROBE, Strengthening the Reporting of Observational Studies in Epidemiology.

*ICMJE Trial Registration Policies*

Nine of the 13 journals endorsed ICMJE trial registration policies and 4 did not (Figure 3). The RCTs published in the 9 journals that endorsed ICMJE registration policies included a trial registration number 67.4% (182/270) of the time. The RCTs published in the other 4 journals included a trial registration number 67.5% (81/120) of the time. No association existed between endorsement of ICMJE and reporting of a trial registry number (OR=1.00, 95% CI: 0.61-1.61).

## Exploratory Outcome

Six of the included oncology journals adhered to REMARK Guidelines (Table 1). No other mention of oncology-specific guidelines was found.

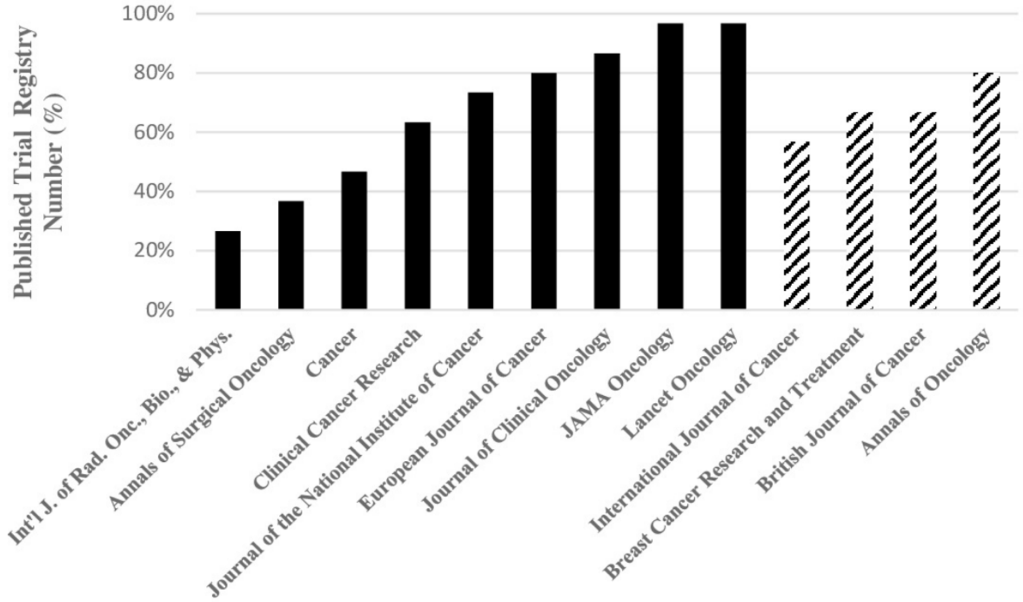

*Figure 3. Histogram of percent of randomly sampled randomized controlled trials (RCT) that included a trial registration number published in journals that adhere to International Committee of Medical Journal Editors (ICMJE) trial registration policies (solid) and those that do not (striped).*

### *Discussion*

Oncology journals support the use of reporting guidelines more often than journals in other medical specialties. Only one journal in our sample did not adhere to any RGs. Recent investigations have found no adherence to RGs in 48% (32/67) of hematology journals[173], 41% (11/27) of emergency medicine[174], and 41% (15/37) of critical care journals.[178]

Specific guidelines such as CONSORT and PRISMA still show a need for greater endorsement since several journals do not mention these guidelines. Evidence suggests that adherence to CONSORT and PRISMA improves some aspects of study methodology in oncology. These studies also called into question important items that remain underreported [179–181]. Additional studies corroborate these findings in other medical specialties [182,183] Specifically, key items such funding source, proper adherence to study protocol, sample size calculation, adverse events, and description of the trial's design have been found to be underreported [162,179,184,185]. The same trends have been observed in oncology SRs; however, these studies have also noted that risk of bias evaluations are infrequently reported [180]

Calls have been made in the past for increased transparency in clinical trial reporting for the sake of patient outcomes and research integrity [186]. RGs were designed to increase scientific transparency and integrity. For example, CONSORT requires trial registration and the reporting of the registration number, which may prevent the selective reporting of outcomes upon publication. A beneficial aspect of RGs for peer reviewers and editors includes the ease by which the methodological rigor of a trial may be evaluated. This is particularly beneficial for junior authors and reviewers who wish to familiarize themselves with aspects of high-quality study designs. The submission of a guideline checklist along with a manuscript may also decrease the time burden for reviewers and editors who choose to investigate the methodological quality of a manuscript.

Our secondary objective was to determine if journal endorsement of a guideline or policy affected the design and reporting of an RCT. Our results demonstrate that the journal adherence to CONSORT guidelines increased the likelihood of authors publishing a CONSORT flow diagram. This finding indicates that oncology journal adherence to CONSORT has a positive effect on reporting practices within oncology trials. And while publication of a participant flow diagram may fail to predict adherence to other CONSORT items, our finding nonetheless demonstrates that good reporting practices are more likely to occur in CONSORT-endorsing oncology journals.

With regard to trial registration, our results demonstrate no association between oncology journal endorsement of ICMJE clinical trial registration policies and author publication of a registry number. The rates of publication of a registry number were similar in ICMJE-endorsing and non-ICMJE-endorsing journals (67.5% and 67.4%, respectively). We recognize a need for improvement, not only because one-third of the analyzed trials did not direct the readership to the registration page, but also because journals that endorsed the ICMJE trial registration policy failed to distinguish themselves from non-endorsing journals. Here, oncology journal policy can help steer clinical trials in a more transparent direction by enforcing the registration of clinical trials and the reporting of the registry number in published manuscripts.

Some areas of medicine, such as oncology, have study designs that are unique and require their own specific guidance with regard to methods and reporting. Therefore, our exploratory outcome was to determine the rates of adherence to oncology-specific RGs. We found that 6 journals adhered to REMARK Guidelines for tumor marker prognostic studies, making it the most popular and only oncology-specific RG within our journal cohort. The rate of adherence to REMARK is encouraging, given the unique study design for which it was created, and the fact that not all journals within our sample accept tumor marker prognostic studies. This finding reflects the overall theme that oncology journals adhere to RGs at a higher rate than journals in other specialties.

For journals that do not currently adhere to RGs or registration policies, a first step may be to simply refer authors to the EQUATOR Network. The EQUATOR Network is the premier clearinghouse for RGs; it was established to aid authors and reviewers in the reporting and evaluation of scientific research, and it is committed to strengthening the integrity of scientific research [187]. The network has produced algorithms to aid researchers who are unfamiliar with RGs to determine which one is most suited for their study design. The network also displays RGs for popular study designs on its homepage to help mitigate the time burden of choosing the correct guideline. Only 3 journals in our sample referred authors to the EQUATOR Network's website.

The EQUATOR Network is currently publicizing and organizing its first project dedicated to a single medical specialty. The EQUATOR Network Oncology Project is designed to increase awareness, address barriers to adherence, and augment the use of RGs in the oncology literature [188]. The future goal of the project is to establish an expert advisory group composed of multiple stakeholders in oncology that share the common goal of using RGs, in part, to increase the quality of oncology research. The EQUATOR Network Oncology Project is currently in the early stages of development, and its present and future plans can be found on the EQUATOR Network's website.

Other methods beyond RGs that are designed to increase research transparency have also been implemented. *The BMJ* and *BMJ Open* both require a declaration of transparency on behalf of all primary authors of clinical trials, with the aim of reducing the incidence of selective reporting bias, which is frequently found in both oncology and hematology journals [36,189]. Additionally, most journals require a declaration of any conflict of interest that the authors may have, with the aim of increasing the integrity and objectivity of published research.

### *Limitations*

One limitation of our analysis with respect to our secondary objective is that the inclusion of a flow diagram may fail to predict a trial's adherence to other CONSORT statement items. Recent studies have demonstrated variable adherence to CONSORT statement items.[190,191] Therefore, our finding that manuscripts in CONSORT-adhering journals more often publish a participant flow diagram may not be generalizable to all CONSORT statement items.

### *Conclusion*

To conclude, RG adherence in oncology journals is better overall compared with other medical specialties that have been investigated, but nonetheless, adherence to individual RGs needs improvement. We have demonstrated that mentioning CONSORT increases the likelihood of author adherence. The benefits of RG adherence have been demonstrated as well as some solutions to potential barriers to uptake and adherence. Ongoing efforts are being made to improve the quality of oncology research, and we encourage support of these efforts, which may begin with a reference to RGs in the instructions for authors page on oncology journal websites.

# CHAPTER IV

# TIDieR CHECKLIST EVALUATION OF CLINICAL TRIAL INTERVENTION REPORTING FOR RECENT FDA-APPROVED ANTICANCER MEDICATIONS

***This work was previously published in the* British Medical Journal: Evidence Based Medicine *with the following citation:***

Wayant C, Bindernagel R, Vassar M. TIDieR checklist evaluation of clinical trial intervention reporting for recent FDA-approved anticancer medications. BMJ Evid Based Med. Published online October 25, 2019. doi:10.1136/bmjebm-2019-111249

---

## *Introduction*

The ability to replicate and effectively implement new healthcare interventions is essential to advancements in patient care. The ability to replicate and implement clinical trial interventions is especially important, since trial methodology lends itself to more trustworthy, replicable results[192]. However, in the past, clinical trials have been shown to exhibit low-quality methods[193,194] and low-quality reporting[195,196], which may compromise a physician's ability to critically appraise trial results and decide whether the intervention is suitable for their practice. Owing to the poor overall methodological and reporting quality of clinical trials, much of the focus for improving clinical trials centered on items such as proper randomization and blinding[197–199]. Much less attention has been dedicated to the reporting of clinical trial interventions[200]. What has conspicuously not been discussed is the actual quality of reporting of clinical trial interventions. Absent high-quality reporting of all aspects of healthcare inventions, implementing interventions effectively becomes difficult. More concerning, however, is that physicians may be incapable of assessing how evidence from highly regimented clinical trials may translate to real-world patient care.

In a recent analysis of cancer chemotherapy trial interventions, it was found that only 11% (30/262) reported all key intervention elements[201]. The most common elements not reported were specific pre-medications and dose-adjustment protocols. Such findings are a cause for concern, especially since newly-approved cancer medications are expensive[202] and may only extend patient survival by a median of 2.1 months[116]. Moreover, many FDA (Food and Drug Administration) oncology drug approvals are now based on non-comparative studies with a primary endpoint that is a surrogate for OS[3,203]. Surrogate endpoints may exaggerate the real-world effectiveness of these novel drugs[204], so the need for clearly and comprehensively described interventions is vital to translating new data to clinical practice.

Given the influence that clearly described healthcare interventions can have on the quality of patient care, and the consequence of poorly described interventions, we aim to evaluate the completeness of intervention reporting in pivotal oncology clinical trials. Our selected cohort of oncology clinical trials consists of trials which formed the basis for novel FDA drug approvals. Our primary research questions are how well-described are these pivotal drug interventions and what are the potential consequences of any identified gaps in reporting.

### *Methods*

We extracted all FDA hematology/oncology drug approvals for anticancer medications from 2017 and 2018 from the FDA website[205]. Using the clinical trial registry identifier included in these approvals or a PubMed search, we matched the approvals to published clinical trials. If the trial registry listed associated publications, we individually reviewed each publication to determine which, if any, contained data that could be matched to the FDA approval summary. If none of the associated publications matched, a PubMed search was used to identify relevant trials. Trials were matched to an FDA approval on the basis of PICO (population, intervention, comparison, outcome), sample size, and reported efficacy data. For example, if the FDA approval summary stated that a phase 3 trial with 200 patients and a primary endpoint of OS was conducted, we searched for trials that matched these criteria, and others if necessary. If no trials were able to be matched, the drug and its approval were excluded. We only included studies of anticancer medications (e.g., solid tumors, leukemia, lymphoma) and excluded any interventions for benign hematologic diseases (e.g., sickle cell anemia).

The matched and included published trials were assessed using the TIDieR (Template for Intervention Description and Replication) checklist[206]. Along with the published trial, all supplemental documents (e.g., protocol, appendices) were investigated. The TIDieR checklist was designed using best-practice, robust Delphi methods by experts in the field of research methods and reporting. The 12 TIDieR checklist items cover the full breadth of intervention reporting: from materials used in the intervention to assessments of treatment compliance and fidelity. A full list of checklist items is available alongside our protocol via the Open Science Framework[207]. Two authors (CW, RB) extracted all data for all TIDieR checklist items for all included published trials using a standardized Google Form created to reflect the 12 TIDieR checklist items. In addition to the TIDieR checklist items, we extracted data related to the design of the trial (e.g., randomization, blinding, study phase) as well as whether a trial protocol was available. We defined a protocol as a document that describes the background, rationale, objectives, design, methodology, statistical considerations, and organization of a clinical research project. Thus, supplemental methods sections were not considered a protocol for the purpose of data analysis, but could be considered when evaluating adherence to TIDieR. After investigation of all published documents with the TIDieR checklist, we reviewed drug labels for each included drug approval. Information gathered from the drug labels did not contribute to TIDieR adherence, since TIDieR is designed to assess the reporting of published trial interventions. Instead, data gathered from the drug label was gathered to determine whether additional information is presented to the FDA that is not presented to the public in a journal article.

We scored each TIDieR item on a three-point Likert scale with the following categories: 0 (not reported), 1 (partially reported), 2 (fully reported). We then summed the scale score for each of the 12 TIDieR items to generate a composite scale score ranging from 0 (poor reporting) to 24 (full reporting) for each included trial. We followed the guidance of Yamato, et al[208], who previously showed that this composite scale score may be used to determine the overall quality of intervention reporting.

The primary outcome of this study was the completeness of reporting of pivotal oncology drug trial interventions that formed the basis for recent FDA approvals. Secondary analyses include evaluating whether clinical trials with accessible protocols exhibited higher-quality reporting and comparing subgroups of trials. Trials were stratified by design, journal, clinical phase, sample size, drug class, and by novelty in class.

For the calculation of summary statistics, medians, and interquartile ranges (IQR) we used Google Sheets. Our percent adherence rates for TIDieR items reflect a denominator of 192, which corresponds to 2 (full TIDieR Item reporting) multiplied by 96 (number of included trials). Thus, a hypothetical item may be partially reported by all clinical trials (each scored as 1 for partial reporting), but the overall percent adherence would be 50% (96/192). To avoid misleading statistical reporting, we report percent adherence rates as well as the number of trials that fully report, partially report, or did not report an item. We prespecified a multiple regression model to investigate the association between journal, funding source, and protocol availability on increased TIDieR composite score of clinical trial intervention reporting. However, our sample of trials aggregated heavily into industry funding and only 3 journals, rendering our multiple regression analysis difficult. Instead, we used the nonparametric *k*-sample test of equality of medians to investigate whether publication of a protocol results in higher median TIDieR composite scores. We used a nonparametric test because our data was not normally distributed upon visual inspection of a histogram and Shapiro-Wilk test of normality. Stata 15.1 was used for all analyses. We prespecified an alpha threshold of .05 for statistical significance.

### *Results*

We identified 121 new listings on the FDA hematology/oncology FDA approvals and safety notifications page[205], of which 36 were excluded (Figure 4). The 85 included listings were all for anticancer medications and were underpinned

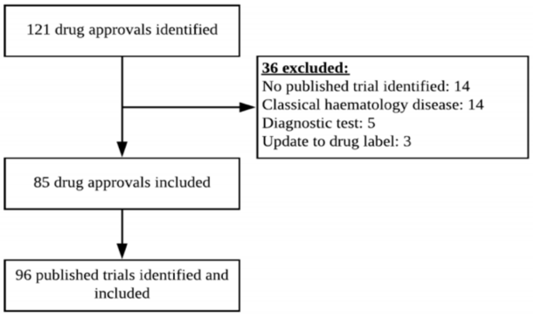
121 drug approvals identified
36 excluded:
No published trial identified: 14
Classical haematology disease: 14
Diagnostic test: 5
Update to drug label: 3
85 drug approvals included
96 published trials identified and included

*Figure 4. Flow diagram of included and excluded studies.*

by a total of 96 clinical trials published in peer-reviewed journals. The 96 clinical trials were most often published in *New England Journal of Medicine* (n = 40, 41.7%), *Lancet: Oncology* (n = 23, 24.0%), and the *Journal of Clinical Oncology* (n = 12, 12.5%). Included trials were most often funded solely by industry (n = 83, 86.5%), followed by mixed funding sources (n = 9 with partial industry, n = 1 without partial industry). Trials were most often randomized (n = 60, 62.5%), open label (n = 68, 70.8%), and phase III (n = 56, 58.3%). See Table 16 for more trial characteristics.

Overall, the median TIDieR composite score was 17 (IQR 2), corresponding to between 8 and 9 (out of 12) fully reported TIDieR items. No trials reported all TIDieR items. Seven TIDieR items had greater than 90% adherence across all clinical trials (Table 2), with 2 items showing 100% adherence: Item 1, name of intervention, and Item 2, rationale for the intervention. Three items were poorly (<5%, each) or moderately (<50%) reported: Item 5, intervention provider (including training and expertise), Item 7, types of institutions where clinical trial was conducted (including infrastructure and relevant features), and Item 11, if and how intervention compliance was assessed. Subgroup comparisons of TIDieR compliance are available via the Open Science Framework[207]. There was no difference between subgroups. The median TIDieR composite score when drug labels were included rose to 18 (IQR 1).

Qualitatively, for the Items 5, 7, 11 that showed poor to moderate adherence (<50%), we identified key action items to improve the quality of intervention reporting. For Item 5, which relates expertise, background and any specific training given to intervention providers, we found a paucity of data. Mostly, protocols would provide a generic statement of "trial personnel" that administered study medications, but failed to mention if these personnel were the investigators, pharmacists, or others. Patient-self administration was scored as partial reporting since there was little detail about training of patients to self-administer. For example, "patients randomized to the duvelisib arm self-administered 25 mg capsules twice daily"[209] exemplifies statements related to patient self-administration. For Item 7, which relates to the description of types of institutions that functioned as trial centers, including infrastructure and relevant features, all clinical trials were scored as "not reported". Despite nearly all clinical trials listing trial centers by name, neither the type of institution, nor the capabilities and infrastructure of these centers were described.

| Item | Full reporting | Partial reporting | No reporting | Overall score (%) |
|---|---|---|---|---|
| 1. Name the intervention. | 96 | 0 | 0 | 192/192 (100%) |
| 2. Rationale for the intervention. | 96 | 0 | 0 | 192/192 (100%) |
| 3. Materials used in intervention (e.g., dose and formulation). | 92 | 4 | 0 | 188/192 (97.9%) |
| 4. Procedures of intervention (e.g., administration procedure). | 91 | 5 | 0 | 187/192 (97.4%) |
| 5. Expertise, background and any specific training given to intervention providers. | 0 | 7 | 89 | 7/192 (3.6%) |
| 6. Mode of delivery of the intervention (e.g., oral/intravenous, alone/group). | 11 | 82 | 0 | 104/192 (54.2%) |
| 7. Types of trial locations, including infrastructure or relevant features. | 0 | 0 | 96 | 0/192 (0.0%) |
| 8. Number of times the intervention was delivered and over what period of time including the number of sessions, their schedule, and their duration, intensity or dose. | 94 | 2 | 0 | 190/192 (99.0%) |
| 9. Plans for individual personalization of treatment (e.g., dose reductions). | 86 | 5 | 5 | 177/192 (92.2%) |
| 10. Study-level modifications to intervention planned (e.g., induction and maintenance). | 60 | 33 | 0 | 153/192 (79.7%) |
| 11. How intervention compliance is assessed. | 11 | 37 | 48 | 59/192 (30.7%) |
| 12. How well was the intervention delivered (e.g., treatment delays, discontinuations). | 93 | 0 | 3 | 186/192 (96.9%) |

*Table 2. Reporting of TIDieR items by included trials (n = 96).*

For Item 11, which relates to whether compliance to the intervention was assessed, we found little information past accountability of study drugs (e.g., accounting for number of pills dispensed and used at follow up). According to the TIDieR checklist[206], assessments of intervention adherence require more than simple assessments of drug quantity consumed — they also relate to how the drugs were consumed. For example, "subjects will keep a daily diary to record dosing compliance, which will also be assessed at each clinic visit by means of a capsule count in the returned bottle. Late doses (i.e., 4 or more hours after scheduled time) should be noted in the diary. Doses that are late by more than 12 hours should be skipped and recorded in the dosing diary as missed."[210] was scored as full reporting; whereas, "the study personnel will account for all investigational products dispensed to and returned from the subject"[211] was scored as partial reporting because it can be inferred that drug quantity accountability was assessed.

The *k*-sample test of equality of medians indicated that the publication of a protocol resulted in significantly higher TIDieR composite scores ($\chi^2$ (1) = 32.0, *P* < .001). Median TIDieR composite score in studies with a protocol (n = 62) was 18 (IQR 1) versus 16 (IQR 1.75) for trials without a protocol (n = 34).

### *Discussion*

The results of this investigation reveal stark differences in the quality of reporting of TIDieR checklist items in pivotal oncology trials. All of the included clinical trials formed the basis for new FDA hematology/oncology drug approvals, and all exhibited strengths and weaknesses, as it relates to the reporting of the interventions. On one hand, all clinical trials listed the drug name and dose and the rationale for the intervention (Items 1 and 2). Included clinical trials also frequently described the number of times the intervention was delivered (including description of treatment cycles) (Item 8), any rules for individual modifications to the intervention (Item 9), and the number of treatment discontinuations or delays (Item 12). Seemingly, all of these items would be expected in an oncology clinical trial, but a previous study of the reporting of oncology interventions suggests that these items cannot always be assumed present[201]. Overall, because the included oncology clinical trials reported items homogeneously — a high proportion did or did not report specific items — we believe that targeted improvements will strengthen all future clinical trials. Namely, we recommend including a description of trial center infrastructure and capabilities (rather than just the name), describing who is administering the interventions (including special training or instructions given), and describing how intervention compliance will be assessed. We further recommend the publication of a study protocol, since better adherence to TIDieR was shown (equivalent to a median of 1 more item, of 12, reported) and inclusion of a protocol should presumably require little effort if one was written for the trial. Last, we recommend that journals scrutinize their reporting requirements for interventions, since FDA drug labels included more information than published reports. All of these recommendations may be enforced by new journal, regulatory agency, or trial registry requirements for intervention reporting.

Many previous studies have noted that cancer outcomes differ based on where patients receive treatment[212–214]. A recent risk-adjusted observational study of Medicare beneficiaries showing a significant reduction in OS for patients treated at community hospitals compared to National Cancer Institute, academic centers, and free-standing cancer hospitals exempt from prospective payment systems[215]. Historically, most clinical trials are conducted at academic centers[216], which may hinder the translation of clinical trial outcomes to community settings, where most cancer patients receive treatment[217,218]. Moreover, there may be selection bias regarding patients enrolled in trials that underpin FDA approvals, since these trials are conducted at academic centers and the patients included may differ from those treated in the community. While it is true that the translation of care from clinical trials to real-world settings and from academic to community centers is multifactorial, one small mechanism to clarify contextual factors that contributed to trial outcomes (and how these outcomes may be expected to translate to other locations) is to describe the infrastructure and capabilities of the trial center. For example, in accordance with the TIDieR checklist recommendations, a clinical trial may report the countries of participating centers, types of hospitals that participated, whether care is publicly or privately funded, volume of hospital activity, or the availability of certain facilities or equipment (as relevant)[206]. Similar to the reporting of trial center infrastructure, explicitly reporting who was involved in administering clinical trial interventions, including any training or prior expertise, may be helpful in translating clinical trial intervention outcomes to the real-world.

Novel cancer immunotherapies, such as CTLA-4, PD-1, and PD-L1 inhibitors, have previously been shown to improve cancer patient survival, but often result in high rates of immune-related adverse events[219,220]. Moreover, it has been reported that immune-related adverse events may continue to progress even after withdrawal of immunotherapy[221]. Despite an understanding of immune-related adverse events, a recent SR noted that the reporting of immune-related adverse events in clinical trials was suboptimal, although better in trials published more recently and in higher impact factor journals, such as those included in our study[219]. Nonetheless, our results indicate that while included trials contained details about how many patients discontinued or delayed the intervention, very little information was given on the methods of assessing compliance. For example, whether outpatient trial participants were required to keep a detailed diary of how many doses were taken, delayed, or missed. Further, the quality of these doses, since many trials include detailed instructions as to how each dose is to be taken (e.g., with food). Given the rate of adverse events, it may be expected for deviations from the protocol to occur beyond dose delays or dose discontinuations. The chief reason for detailed compliance information is to determine whether or not treatment effects are likely to be affected by how well patients complied to study procedures, and also to show readers how compliance was encouraged. We recommend future trial authors detail how compliance is to be assessed for the intervention, rather than just the rates of adverse events.

This study has several strengths and limitations. For strengths, we used double data extraction to mitigate bias in the data collection process. We further restricted our analysis to high-profile oncology clinical trials published, and thus our cohort of studies may represent the highest level of reporting of oncology interventions. However, because our sample represents a very unique cohort of oncology clinical trials, the results of this investigation may not be generalizable. The trials in our sample were remarkably homogeneous — TIDieR items were either reported or not by almost all studies. Thus, our recommendations to improve the quality of reporting of oncology clinical trials may require validation in a different cohort of oncology trials. Last, this study was not designed to incorporate oncologist beliefs and attitudes regarding deficits in TIDieR reporting.

In conclusion, we found that key TIDieR items were reported in nearly all included clinical trial interventions (each of which formed the basis for an FDA drug approval). These highly-reported items include the dose and treatment regimen for the intervention and decision rules for individual treatment modifications. However, key items were under-reported that may be useful to oncologists in the community, such as rates of compliance, the infrastructure of the participating trial centers, and the training or expertise for intervention providers. We recommend journals and regulatory agencies adopt and enforce the reporting of all pertinent aspects of clinical trial interventions. Trial registration sites, such as clinicaltrials.gov, may further be restructured to request information required by the TIDieR checklist. We further recommend adherence to TIDieR and the publication of a complete study protocol, since the inclusion of a protocol was significantly associated with increased quality of intervention reporting.

# CHAPTER V

# FINANCIAL CONFLICTS OF INTEREST AMONG ONCOLOGIST AUTHORS OF REPORTS OF CLINICAL DRUG TRIALS

***An abbreviated version of this work was previously published in* JAMA Oncology *as a research letter with the following citation:***

Wayant C, Turner E, Meyer C, Sinnett P, Vassar M. Financial Conflicts of Interest Among Oncologist Authors of Reports of Clinical Drug Trials. JAMA Oncol. Published online August 30, 2018. doi:10.1001/jamaoncol.2018.3738

***This chapter includes the full-length article.***

---

## *Introduction*

In medicine, a conflict of interest may affect patient care in many ways.[222] In clinical care, a physician's receipt of payments from a pharmaceutical company may interfere with one's ability to objectively treat patients. Even small payments, such as meals, increase the likelihood that a physician will prescribe a company's drugs.[223] In clinical research, conflicts of interest can undermine the legitimacy of data reporting. A recent Cochrane SR found that industry-sponsored trials were more likely to report favorable efficacy results and present conclusions that were inconsistent with the study's results than non-industry funded trials.[224] Conflicts of interest, thus, compromise the quality of patient care and public trust in the medical profession.

Concerns about physician financial conflicts of interest (FCOI) led to the passage of the Physician Payments Sunshine Act and the creation of the Open Payments Database (hereafter referred to as Open Payments).[225,226] Open Payments aims to increase the transparency of financial relationships between physicians and industry by making all pharmaceutical industry payments to physicians of $10 or more publicly available. A recent investigation using 2015 Open Payments data found that approximately 48% of United States physicians received payments totaling $2.4 billion from pharmaceutical companies.[227] In oncology, physician FCOI disclosures are particularly important because of the need to bring lifesaving drugs to market while simultaneously ensuring their efficacy and safety. Previous oncology investigations have shown that FCOI disclosures may be inadequate among clinical trialists,[228] editors,[229] and CPG authors[153]; however, a focused investigation of FCOI for clinical trialists who undertook the trials that led to FDA drug approval is warranted. These trials represent the driving force in oncology and rapidly change the trajectory of cancer care. These trials also generate high impact factor publications, prestige for authors, revenue for the pharmaceutical companies, and newsworthy headlines.

Therefore, the primary objective of this cross-sectional analysis of FDA-approved oncology drug trials is to quantify the frequency and amount of industry-author financial relationships. The secondary analysis is to identify the frequency of undisclosed author FCOIs and determine if an increase in General payments resulted in the year following publication. We conducted pre-planned subgroup analyses, stratifying our data by journal and industry sponsor.

### *Methods*

This study was not subject to institutional review board oversight because it did not meet the regulatory definition of human subject research as defined in 45 CFR 46.102(d) and (f) of the Department of Health and Human Services' Code of Federal Regulations.[230]

One of us (CW) searched the FDA Hematology/Oncology (Cancer) Approvals & Safety Notifications web page for oncology drug approvals between January 1, 2016 and August 30, 2017 (the start date of our study). The focus of this investigation was to assess the FCOIs for authors of drug trials for malignant diseases. We excluded any trials of drugs for benign diseases (e.g., sickle cell disease) and any trial of a diagnostic tool/test. FDA approvals cite clinical trial number(s) rather than a published report; therefore, we identified the published manuscript with the endpoints that formed the basis of each FDA approval. If the FDA approval did not report the clinical trial number(s), we searched the press releases for the pharmaceutical companies that sponsored the drug trial for the clinical trial number(s). If the press releases did not report the clinical trial number(s), we searched ClinicalTrials.gov using PICO (Population, Intervention, Comparison, Outcome) keywords. The process by which we identified clinical trial numbers for drug approvals is detailed in Figure 5. If an FDA approval was based on the pooled analysis of multiple clinical trials, we included all underlying trials.

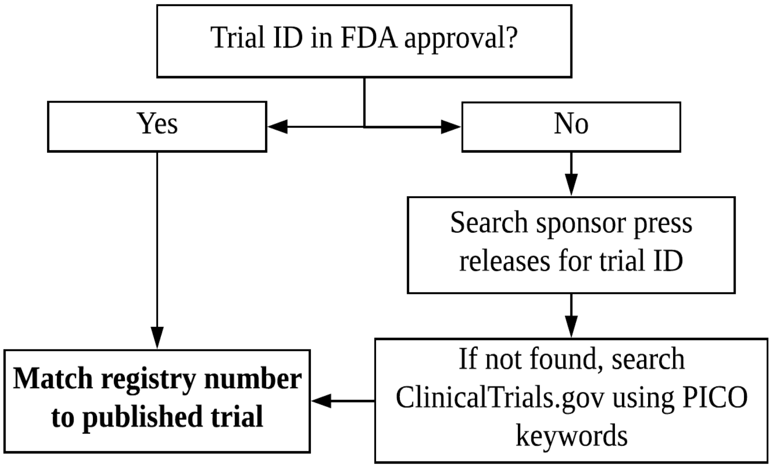
Trial ID in FDA approval?
Yes
No
Search sponsor press releases for trial ID
If not found, search ClinicalTrials.gov using PICO keywords
Match registry number to published trial

*Figure 5. Process of identifying published clinical trials.*

After identifying the published reports, we only included those that were published after September 2013 which corresponds with the earliest month and year of payments catalogued in Open Payments. From the published reports, one of us (CW) extracted the following items: title, journal, date of trial registration (considered the start of the financial relationship for that trial), date of publication (considered the end of the financial relationship for that trial), drug for which the approval was based, industry sponsor, author names, author affiliations, and author disclosure statements. Only United States physicians (MD or DO) were included in this analysis, since other degrees (e.g., PhD) and authors from other countries are not catalogued in Open Payments.

Three of us (CW, CM, and PS) then proceeded to search the Open Payments Database for each author’s FCOIs. CM and PS extracted all data first, then CW validated the data by extracting data a second time for comparison. CW was blinded to the data extracted by CM and PS. We used a combination of author name, location of institution, and medical specialty to correctly identify authors. From the Open Payments Database, we extracted all payments from the industry sponsor starting with the year of trial start date to the year of publication, and recorded any time authors disputed a payment. Any disputed payments were subtracted from the final payment amount. We also extracted data from any subsidiaries of a pharmaceutical company. We chose to do so because the parent company stands to profit from a successful subsidiary-sponsored trial. Additionally, we extracted the General Payments (minus food/beverage payments) for the year following trial publication to examine the continuation of the industry-author financial relationship after publication of a high-impact trial.

The Open Payments Database categorized payments into four categories:

1) *General*: These include consulting fees, speaking fees, honoraria, gifts, entertainment, food and beverage, travel and lodging, and education.
2) *Research payments*: Payments associated with a research study, including basic and applied research, and product development.
3) *Associated research payments*: Funding for a research project or study where the physician is named as a principal investigator.
4) *Ownership*: Ownership and investment interest in companies, which describes both the actual dollar amount invested and the value of the ownership or investment interest.

We cross-referenced author disclosure statements from the published report with the payments received from the industry sponsor of the drug. A disclosure statement was considered inaccurate if an author did not completely disclose the financial relationship depicted in Open Payments. For example, if an author disclosed only grant funding from the sponsor of their trial, but received speaking, consulting, or honoraria fees. Authors commonly reported only "personal fees" and we were unable to uniformly determine if this referred to grants, research payments, or personal payments. Therefore, as long as the personal fees were from the sponsoring company, we were forced to consider this sufficient. We did not encounter any Ownership payments; therefore, the Results and Tables exclude this category.

Sums, means, and medians, were calculated using Microsoft Excel. We made violin plots using RStudio and the package ggplot2 for visual representation of the distribution of payments for the three included categories.

***Results***

General Characteristics

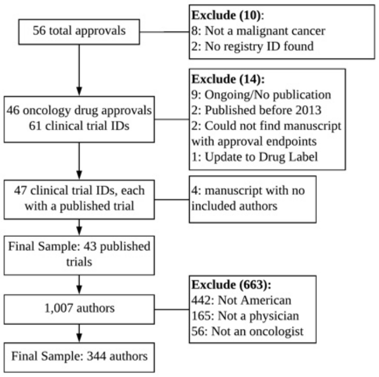
56 total approvals
Exclude (10):
8: Not a malignant cancer
2: No registry ID found
46 oncology drug approvals
61 clinical trial IDs
Exclude (14):
9: Ongoing/No publication
2: Published before 2013
2: Could not find manuscript with approval endpoints
1: Update to Drug Label
47 clinical trial IDs, each with a published trial
4: manuscript with no included authors
Final Sample: 43 published trials
1,007 authors
Exclude (663):
442: Not American
165: Not a physician
56: Not an oncologist
Final Sample: 344 authors

*Figure 6. Flow diagram of included clinical trials.*

Between January 1, 2016 and August 30, 2017, we identified 56 FDA hematology/oncology approvals. Ten approvals were excluded from this analysis. The remaining 46 approvals were based on 61 clinical trial numbers, of which 43 were included for analysis (for exclusions see Figure 6). These 43 clinical trial numbers were each successfully linked with a published trial. In all, 1,007 authors were included. From these 1,007 total authors we identified 344 United States physicians to be included for our primary analysis (Figure 6). There was a median of 11 United States physicians per manuscript (IQR 7.5 - 20). All manuscripts identified were published in one of six journals: *JAMA Oncology, Journal of Clinical Oncology (JCO), New England Journal of Medicine (NEJM), The Lancet, The Lancet: Oncology, The Lancet: Haematology.* All six journals adhere to the International Committee of Medical Journal Editors (ICMJE) policy regarding FCOI and require authors to fill out a disclosure form prior to manuscript submission.

Primary Objective

Of 344 authors, 263 (76.5%) accepted at least one payment, 196 (57.0%) accepted more than $100,000, 48 (14.0%) accepted more than $1,000,000, eight (2.3%) accepted more than $5,000,000, and two (0.6%) accepted more than $10,000,000. The largest amount of money received during the course of a clinical trial was $25,661,335 from Novartis. The median total payments to authors was $195,321 (IQR $532 - $597,628) with a cumulative total of $216,627,353.

For General payments, authors accepted a median of $2,828 (IQR $0 - $19,628), and a total of $6,318,031. For Research payments, authors accepted a total of $513,885 (median of $0). For Associated Research payments, authors received a median of $164,644 (IQR $0 - $551,926) and a total of $209,795,437 (Table 3).

|  |  | **General Payments (minus food)** | **Research Payments** | **Associated Research Payments** |
|---|---|---|---|---|
| **Total Payments** | Median (IQR) | $2,828 ($0 - $19,628) | 0 ($0 - $0) | $164,644 ($0 - $551,926) |
|  | Mean (SD) | $18,336 ($107,087) | $1,494 ($10,841) | $609,870 ($1,843,467) |
|  | Sum | $6,318,031 | $513,885 | $209,795,437 |
| **Disclosed FCOIs** | Median (IQR) | $1,170 ($0 - $20,506) | $0 ($0 - $0) | $81,591 ($0 - $518,546) |
|  | Mean (SD) | $19,544 ($128,467) | $1,173 ($5,500) | $563,049 ($1,900,634) |
|  | Sum | $4,573,269 | $273,207 | $131,753,549 |
| **Undisclosed FCOIs** | Median (IQR) | $3,783 ($58 - $18,793) | $0 ($0 - $0) | $292,273 ($44,909 - $667,547) |
|  | Mean (SD) | $15,861 ($28,249) | $2,188 ($17,457) | $709,472 ($1,719,681) |
|  | Sum | $1,744,762 | $240,678 | $78,041,888 |

*Table 3. Payments to included oncologist-authors (n = 344).*

## Secondary Objective

Of 344 authors, 110 (31.9%) did not fully disclose their FCOI from the trial's sponsor. Of these 110 authors, 79 (71.8%) accepted more than $100,000, 17 (15.5%) accepted more than $1,000,000, and 4 (3.6%) accepted more than $5,000,000. The greatest amount of undisclosed payments accepted during one clinical trial was $15,363,234 from Novartis. For General payments, non-disclosing authors accepted a median of $3,783 (IQR $58 - $18,793), and a total of $1,741,186. For Research payments, non-disclosing authors accepted a total of $240,678 (median $0, IQR $0 - $0). For Associated Research payments, non-disclosing authors accepted a median of $292,273 (IQR $44,909 - $667,547), and a total of $78,041,888.

Twenty-nine authors published trials between 2013 and 2015. These authors were assessed for an increase in General payments (minus food/beverage) the year following publication of their manuscripts. Thirteen (44.8%) received more General payments (minus food/beverage) in the year following than their average amount received in the years previous. The mean year-after increase was $7,770. Thirteen (44.8%) received no payments during and after the completion of their trial. Three (10.4%) received less than their previous year(s) average. The mean year-after decrease for these three authors was $1,585.

*Sponsors*

The 43 trials reported on 23 unique oncology drugs and were published with financial support from 19 unique sponsors (or combinations, thereof). Merck (n = 70), Hoffman-La Roche/Genentech (n = 49), and Bristol-Myers Squibb (n = 45) funded the most authors. The sponsors with the highest proportion of authors with undisclosed FCOI were Pfizer (9/16, 56.3%), Eisai (6/12, 50.0%), and Tesaro (3/6, 50%). Data for each sponsor and median payments to each author by individual drug are available in the Data Supplement.

*Journals*

Authors published most often in *The Lancet: Oncology* (n = 115), *New England Journal of Medicine* (n = 100), and *The Lancet* (n = 70). The journals with the highest proportion of authors with undisclosed FCOI were *New England Journal of Medicine* (46/100, 46.0%), *The Lancet* (26/70, 37.1%), and *The Lancet: Haematology* (11/30, 36.7%). Authors published in *New England Journal of Medicine* had the highest median undisclosed General payments, while authors in *The Lancet* had the highest median Associated Research payments.

### Discussion

Our study found that approximately one in three oncology authors failed to adequately disclose industry financial conflicts of interest (FCOIs). Moreover, the median undisclosed payments exceeded the median disclosed payments for General and Research payments. Further, we demonstrated that physicians who published high-profile oncology drug trials received more General Payments (e.g., honoraria, consulting fees, travel/lodging) in the year following publication. These findings are consistent with what has been shown in the oncology literature.[47,153,231]

Historically, opinions regarding FCOIs in medical research have differed. Most parties would agree that the relationship between industry and physicians has benefited patients with cancer through the development of drugs that improve survival and quality of life. At the same time, there is widespread concern that the pharmaceutical industry's role in medical practice unduly influences professional behavior and judgement.[152,232]

These concerns are likely to grow since industry continues to fund an increasingly large proportion of medical research.[233]

Some skeptics question whether the emphasis on FCOI disclosures misses the mark of preventing bias, arguing that it is the receipt of payments that increases the risk for bias.[234] To them, there is less value in requiring FCOI disclosures, since disclosure cannot retroactively prevent bias. The might also question whether all payments from industry should be called "conflicts", since aligning with industry represents a "confluence" of physician interests to better serve patients.[235]

Countering such skepticism, there is credible evidence that industry-sponsored studies are more likely to report favorable efficacy results[224] and that oncologists who receive industry payments are more likely to prescribe industry drugs.[92] Further, proponents of FCOI disclosure argue that a lack of disclosure can signal potential bias, and allow the reader an opportunity to draw conclusions regarding the validity of the study results. They would further argue that simply disclosing payments is not enough, and that stricter standards and regulations should be applied to industry-author financial relationships.[232]

At minimum, these stricter standards and tighter regulations would include cross-referencing an author's voluntary FCOI disclosure with Open Payments data. Open Payments Database is organized to make searching for authors and their payments easy and quick. To explain the utility of this practice, consider a scenario from our dataset. A physician authored three trials for pembrolizumab: one for the treatment of gastric cancer; two for the treatment of head and neck squamous cell carcinoma. In the author's first publication he declared no competing FCOIs. In his next two publications he reported a financial relationship with Merck. He received payments in years prior to his first publication, all the way through his last publication.

Scenarios like this example were common in our analysis and suggest that authors are willing to disclose FCOIs, but occasionally do not for unknown reasons. Moreover, it may reflect inertia in conforming to standards for FCOI disclosure. However, as also shown in the example, the inertia can be overcome, since he did eventually disclose his FCOIs. The original causes for undisclosed FCOIs must be identified and addressed to continue progress towards complete transparency in FCOI disclosures. Ignoring the barriers to FCOI disclosure is not an option, since financial bias can affect the conduct of clinical trials.[236]

We recommend that journals, trial sponsors, and authors review the Open Payments Database for each clinical trial they conduct or publish. Though, for these parties to commit to this practice, the database must be recognized as accurate. Anecdotally, some have recently given accounts of discrepancies in the database, but these accounts focus on misattributed General Payments, such as meals.[237] Furthermore, when larger issues arose soon after the creation of the database, the Centers for Medicare and Medicaid Services took drastic measures to resolve and prevent future errors,[238,239] making it unlikely that a six-figure research payment to an author could be misattributed. Such large payments were common among oncologists in our analysis. The current nature of oncology research relies on expert opinion and experts who frequently work alongside industry in clinical research, and one would expect such relationships to have a financial component.[240] These payments do not, and should not, prevent the publication of a clinical trial; but, failure to disclose these payments raises suspicion for bias.

As the pharmaceutical industry's involvement in medical research has grown, the medical community has directed increasing attention to the issue of financial bias.[41,241] Now, complete FCOI disclosure is often seen as a minimum obligation. In light of this investigation's findings, a critical next question is: which parties, if any, are responsible for ensuring complete disclosures of FCOIs? Currently, physicians, industry, and journals seem to work separately in handling the disclosure of FCOIs, but it appears to us that complete disclosure will require these parties to work in concert.

Trialists must learn to prioritize disclosure complete disclosure of FCOIs so that it becomes second nature, akin to recruiting participants. Pharmaceutical companies should require authors they fund to completely disclose all FCOIs in all trial-related publications. Such a policy could be emphasized prior to entering a financial relationship and include this requirement in any mutually approved contract. Journals should act to educate and require peer reviewers to cross-reference an author's voluntary FCOI disclosure with Open Payments data. Cross-referencing author disclosures and payments could be expedited by requiring authors to submit a link to their Open Payments information. Then, any discrepancies could be included in peer reviewer comments to authors and authors could be required to respond. If all three parties buy in, undisclosed FCOIs could be eliminated either by completely disclosing them or clarifying any rare instance of a misattributed payment.

The limitations of our study include potential inaccuracies of the Open Payments Database and human errors in data extraction. Even though most physicians are listed with middle initials, specialty, and location, not all physicians have this information. In such cases, it is possible that we were led to extract data from the wrong physician. All efforts were taken to mitigate this possibility, including data extraction and verification from three authors.

To conclude, we found that close to one in three physician authors of FDA-approved drug trials failed to completely disclose FCOIs. Three parties — authors, sponsors, and journals — share in the responsibility for correcting this problem. Each plays a key role in addressing concerns for financial bias in high-impact oncology clinical trials. We argue that, so long as industry plays a role in funding oncology clinical trials, complete physician disclosure of FCOIs should be considered a minimum obligation.

# CHAPTER VI

# EVALUATION OF SPIN IN ONCOLOGY CLINICAL TRIALS

***This work was previously published in* Critical Reviews in Oncology/Hematology *with the following citation:***

Wayant C, Margalski D, Vaughn K, Vassar M. Evaluation of spin in oncology clinical trials. Crit Rev Oncol Hematol. 2019;144:102821.

---

### *Introduction*

When authors misrepresent, distort, or otherwise selectively feature specific research data they introduce “spin” to the literature. The prevalence of spin has been quantified in a recent SR, which found that a median of 56.8% of trials contain some form of spin[242]. The effect of spin has been demonstrated in a two-arm, parallel group randomized trial involving 300 oncologists who were asked to evaluate a trial abstract with a nonsignificant primary endpoint[52]. Half were assigned to read an abstract with an overly optimistic conclusion about the intervention, while the other half read an abstract that concluded no benefit of the intervention. Oncologists who received the abstract with the overly optimistic conclusion rated the intervention as more effective, the trial as less rigorous, and were more likely to read the full text of the trial.

In 2016, a trial was published that examined the effect of adjuvant sunitinib on advanced stage renal cell carcinoma post nephrectomy. This trial, which formed the basis for FDA approval, used improvements in a surrogate endpoint (disease-free survival) as evidence of drug efficacy, and relegated the collinear Kaplan-Meier curves for OS to the Supplement.[243] Despite OS being the secondary endpoint, the eight year follow-up data and hazard ratio of 1.01 (95% CI, 0.72 to 1.44) was highly suggestive of no survival benefit. Indeed, a letter in reply to this trial was written to emphasize that the goal of cancer therapy is to extend survival and mitigate adverse events, neither of which was shown in this trial of sunitinib[244]. Some have described the FDA approval of sunitinib as "regulatory capture"[245], which occurs when parties with high-stakes interest in a policy decision overpower other parties with less interest to achieve an intended outcome. We have an additional concern: how certain research data can be highlighted or omitted in order to shift perceptions of a drug's efficacy. In this case, by not mentioning the nonsignificant OS data in the abstract and placing the Kaplan-Meier graph in the Supplement, the authors of the sunitinib study framed their printed study around a statistically significant surrogate endpoint and omitted visual evidence of the nonsignificant survival benefit.

There are many forms of spin in the reporting of medical research findings, but at its core spin is an attempt to beautify or omit unfavorable results.[242] Abstracts may be most susceptible to spin because of the limited word counts enforced by journals. Further, the consequences of spin in abstracts may be more severe. There is evidence that many physicians only read the abstract of most research articles[246–248]. Moreover, institutions in low- or middle-income countries may not have the resources to access the full text of articles. Thus, abstracts must be accurate synopses of full manuscripts and avoid misleading conclusions about drug efficacy. In oncology, spin occurs in the abstracts as well as the full text of manuscripts[185,249]. Trial authors may omit toxicity results or selectively report endpoints based on statistical significance. And while all areas of medicine including oncology are susceptible to spin, we argue that the oncology literature may be most susceptible due, in part, to the presence of surrogate endpoints that are designed to predict clinical benefits to patients.

Surrogate endpoints are often acceptable in oncology trials[250], even for FDA approval.[3] However, OS is considered the ideal survival endpoint in oncology trials owing to its objectivity and relevance to patients.[251] But, OS requires increased sample size and follow-up duration, which may delay the approval or development of new therapies.[252] Owing to these factors, surrogate endpoints have increased in popularity as primary endpoints since they often require fewer patients and less time to measure.[250] Despite their popularity, surrogate endpoints are often imperfect measures of OS and frequently have larger effect sizes.[253] The fact that surrogate endpoints are often imperfect measures of treatment effectiveness, they may show discordant results and OS may be required to clarify treatment utility in practice. However, because OS is more often statistically nonsignificant, authors may be tempted to spin toward the surrogate endpoints that tend to have larger effect sizes.

Regardless of which endpoint is primary and which is secondary, selectively emphasizing a secondary endpoint or subgroup analysis means authors are emphasizing fragile, underpowered results. Since surrogate endpoints and OS are almost equally acceptable in clinical trials, oncologist-authors may feel comfortable focusing on whichever of these endpoints is statistically significant. Therefore, the primary objective of this novel investigation is to evaluate the frequency and manifestations of spin in oncology clinical trials that measured a surrogate endpoint and OS.

### *Methods*

We searched PubMed on March 30, 2018 to identify clinical trials published in 2017 reporting at least one key surrogate endpoint and OS published in ten key journals. The exact search strategy is publicly available via the Open Science Framework.[254] The six key surrogate endpoints were PFS, disease-free survival, objective response rate, complete response, time to progression, and time to treatment failure. These surrogate endpoints were selected based on the Food and Drug Administration's (FDA) "Guidance for Industry: Clinical Trial Endpoints for the Approval of Cancer Drugs and Biologics" document.[255] The following journals were searched: *Journal of Clinical Oncology*, *The Lancet: Oncology*, *JAMA Oncology*, *Cancer*, *Annals of Oncology*, *Journal of the National Cancer Institute*, *British Journal of Cancer*, *European Journal of Cancer*, *New England Journal of Medicine*, and *The Lancet*. Search results were added to a PubMed collection and uploaded to Rayyan[256].

One of us (CW) screened all articles for inclusion. To be included an article had to meet the following criteria: randomized clinical trial with a head to head comparison, measure both a surrogate endpoint from the FDA "Guidance to Industry" list and OS, and conduct a superiority analysis. We excluded articles that were not clinical trials, clinical trials with an incompatible design (e.g., cluster, crossover, single arm), pooled analyses, noninferiority analyses, and trials of non-oncologic interventions published in the included general medical journals. Three of us (CW, DM, and KV) extracted data for all included trials.

*Definition of Spin*

Our definition of spin was based on Boutron, et al.,[137] which states that spin is the "use of specific reporting strategies, from whatever motive, to 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." We modified this definition to include all trials regardless of the statistical significance of the primary endpoint. Doing so allowed us to capture evidence of spin in abstract conclusions when a surrogate endpoint (primary endpoint) was statistically significant and OS (secondary endpoint) was nonsignificant.

*Objectives*

Our primary objective was to assess the frequency and manifestations of spin in oncology clinical trials that report both a surrogate endpoint and OS. Spin was assessed within a trial (i.e., emphasizing a subgroup analysis when the primary analysis is nonsignificant) and outside a trial (i.e., selective outcome reporting bias - adding, subtracting, or changing the order of trial endpoints compared to a trial registry before publication). In the former (within a trial), primary trial endpoints would match the registry, but they would be reported out of order or with different emphasis. In the latter (selective outcome reporting bias), trial endpoints would not match the registry. Our secondary objective is to compare trials with OS as the primary endpoint and with OS as the secondary endpoint.

*Spin in the abstract title and results*

We considered there to be evidence of spin if a study title suggested treatment effectiveness where none exists. For example, if a title began with "First line use of…", despite showing no significant benefit of the intervention, this may be spin and mislead readers about the study conclusions. We considered there to be spin in abstract results when a trial emphasized statistically significant results out of order (e.g. subgroup before overall analysis, secondary endpoint before primary endpoint), reported a per-protocol analysis when intention-to-treat was prespecified, or used “trend statements” in the description of statistical significance (e.g., “trend toward significance”). We did not consider it to be spin if a trial omitted a secondary endpoint (including OS) from the abstract results section, since this could be interpreted as the standard reporting of results.

### *Spin in the abstract conclusions*

We considered there to be evidence of spin when a trial interpreted statistically nonsignificant results as showing treatment equivalence or comparable effectiveness, focused on a significant subgroup or within-group comparison, used unjustified, optimistic statements in the description of outcomes, emphasized subgroups or modified treatment populations, distracted from nonsignificant findings by stating that the nonsignificant results were due a trial design issue (e.g., underpowered), or claimed treatment benefit from a statistically significant surrogate endpoint when OS was nonsignificant. We considered using only a surrogate endpoint as evidence of treatment benefit as spin because surrogate endpoints have been shown to be poor predictors of OS.[61] Abstract conclusions frequently state that the primary, surrogate endpoint was met while maintaining a focus on patient-important endpoints, such as OS. A clear statement that OS data was nonsignificant was not required. In phase 2 trials or interim analyses of phase 3 trials, we did not consider there to be spin if authors stated that further investigation was necessary to confirm the present findings.

### *Statistical Analysis*

Summary statistics (frequencies and proportions) were calculated using Google Sheets. We used Stata 15.1 (Stata Corp, LLC; College Station, TX) and Fisher’s exact test to compare differences in categorical endpoints.

### *Results*

Of the 620 articles retrieved, 124 were included. Articles were excluded mostly for not being oncology trials published in general medical journals, being nonrandomized (including single-arm trials), or for not measuring both a surrogate from the FDA guidance document and OS. Figure 7 itemizes all exclusions.

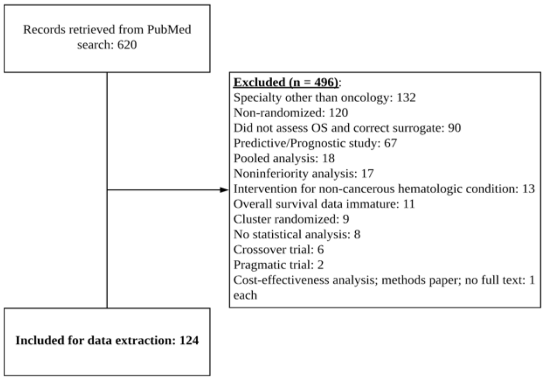
Records retrieved from PubMed search: 620
Excluded (n = 496):
Specialty other than oncology: 132
Non-randomized: 120
Did not assess OS and correct surrogate: 90
Predictive/Prognostic study: 67
Pooled analysis: 18
Noninferiority analysis: 17
Intervention for non-cancerous hematologic condition: 13
Overall survival data immature: 11
Cluster randomized: 9
No statistical analysis: 8
Crossover trial: 6
Pragmatic trial: 2
Cost-effectiveness analysis; methods paper; no full text: 1 each
Included for data extraction: 124

*Figure 7. Flow diagram of included and excluded studies.*

Table 4 itemizes characteristics of the included trials and describes the proportion of trials with spin associated with each characteristic. Overall, in the 124 trials there were 126 primary endpoints: 71 were surrogate endpoints and 55 were OS. In five trials, OS and a surrogate endpoint were co-primary endpoints, while in nine trials the surrogate endpoint and OS were co-secondary endpoints. The most common surrogate endpoints measured were PFS (n = 46), followed by disease-free survival (n = 17), and overall response rate (n = 3). The primary endpoint was clearly described in 86.3% (107/124) of abstracts. In the majority of cases, the primary endpoint was not statistically significant (79/126; 62.7%, 95%CI 54.0%-70.7%).

We found evidence of spin in 46 of 124 (37.1%, 95%CI 29.1%-45.9%) of trial abstracts (Table 5). There was no evidence of spin in trial titles. Spin was present in 19 (15.3%, 95%CI 9.9%-22.4%) abstract results and 40 (32.3%, 95%CI 24.7%-40.9%) abstract conclusions. Of the 118 trials that reported a trial registration number, 10 (8.5%, 95%CI 4.7%-14.9%) selectively reported their endpoints. Sixteen (12.9%, 95%CI 8.1%-19.9%) RCTs had spin in both the abstract results and conclusions.

When OS was a primary endpoint there was evidence of spin 29.1% (16/55; 95%CI 18.8%-42.1%) of the time. OS was statistically significant in 3 of these trials; however, evidence of spin in each trial was due to selective reporting bias, which indicates that authors deviated from the trial protocol to report the statistically significant OS data as the primary endpoint.

| Characteristic | | Total No. (%) | No. With Spin (%) |
|---|---|---|---|
| *Primary endpoint* | | | |
| | Overall survival | 46 (37.1) | 13 (28.2) |
| | Surrogate endpoint | 60 (48.4) | 24 (52.2) |
| | Both | 9 (7.3) | 3 (6.5) |
| | Neither (e.g., both secondary endpoints) | 9 (7.3) | 6 (13.0) |
| *Journal* | | | |
| | Journal of Clinical Oncology | 31 (25.0) | 16 (34.8) |
| | Lancet Oncology | 30 (24.2) | 5 (10.9) |
| | Annals of Oncology | 19 (15.3) | 10 (21.7) |
| | New England Journal of Medicine | 16 (12.9) | 4 (8.7) |
| | JAMA Oncology | 9 (7.3) | 3 (6.5) |
| | British Journal of Cancer | 8 (6.5) | 5 (10.9) |
| | Cancer | 6 (4.8) | 3 (6.5) |
| | Lancet | 4 (3.2) | 0 (0.0) |
| | Journal of the National Cancer Institute | 1 (0.8) | 0 (0.0) |
| *Control arm* | | | |
| | Active drug only | 85 (68.5) | 31 (67.4) |
| | Active + Placebo | 22 (17.7) | 4 (8.7) |
| | Placebo | 12 (9.7) | 9 (19.6) |
| | Other | 4 (3.2) | 1 (2.2) |
| | Surgery | 1 (0.8) | 1 (2.2) |
| *Funding source* | | | |
| | Industry | 68 (54.8) | 22 (47.8) |
| | Mixed (with Industry) | 25 (20.2) | 13 (28.3) |
| | Public (e.g., government) | 24 (19.4) | 6 (13.0) |
| | Private (e.g., foundation) | 4 (3.2) | 2 (4.3) |
| | Hospital | 1 (0.8) | 1 (2.2) |
| | Not mentioned | 1 (0.8) | 1 (2.2) |
| | Other | 1 (0.8) | 1 (2.2) |

*Table 4. Characteristics of included studies and the proportion of studies with those characteristics that contained spin.*

Thus, when OS was the primary endpoint, spin was primarily used to distract from the nonsignificant OS data. When OS was a secondary endpoint there was evidence of spin 43.5% (30/69; 95%CI 32.4%-55.2%) of the time. OS was statistically significant in 2 of these trials, meaning authors were most likely to frame their conclusions around statistically significant surrogate endpoint data, rather than patient important outcomes. There was no significant difference in the rates of spin when OS was a primary or secondary endpoint (p = .13), indicating that evidence of spin was used for different reasons depending on which endpoint is primary and which is statistically significant.

| **Location and type of spin** | **No. (%)** |
|---|---|
| *Total abstracts with evidence of spin* | 46 (37.1) |
| *Abstracts with spin in the title* | 0 (0.0) |
| *Abstracts with spin in the results** | 19 (15.3) |
| Focus on statistically significant subgroup analysis | 6 |
| Use of suggestive language (e.g., trend toward significance) | 5 |
| Omit statistically nonsignificant OS primary endpoint data | 4 |
| Focus on hazard ratio, omit confidence interval and p-value | 2 |
| Focus on statistical significance, ignoring small effect size | 1 |
| Other | 5 |
| *Abstracts with spin in the conclusions** | 40 (32.3) |
| Recommend use of drug based on surrogate endpoint alone | 17 |
| Emphasis on statistically significant subgroup analysis | 5 |
| Interpreting a nonsignificant *P* value as showing noninferiority | 5 |
| Focus on flaws in trial design rather than nonsignificant results | 4 |
| Use of suggestive language (e.g., trend toward significance) | 3 |
| Focus on statistical significance, ignoring small effect size | 1 |
| Other | 9 |

*Table 5. Location, type, and frequency of spin in abstracts. Sums may exceed the total because some abstracts contained multiple types of spin in multiple locations.*

Spin in the abstract results was most often due to authors emphasizing a statistically significant subgroup analysis (n = 6). Five trials used rhetoric to spin their data, 4 emphasized a statistically significant secondary endpoint, and 2 reported only a ratio of events that favored the intervention and omitted the confidence interval or p-value that would have shown that the intervention effect was not statistically significant.

Spin in the abstract conclusions was most often due to authors using a statistically significant surrogate endpoint to highlight the efficacy of their intervention, without caution because of nonsignificant OS data (n = 17). All of the included trials had mature OS data. Five trials emphasized a statistically significant subgroup analysis, 5 wrongly interpreted a nonsignificant p-value as showing equivalence between the experimental and control groups, and 4 distracted from nonsignificant findings by critiquing their trial design. Nine trials were classified as having “Other” evidence of spin, and in these cases the authors either claimed that the intervention was beneficial, despite reporting no significant endpoints, or claimed to accomplish another objective that was not established *a priori* (ex., they conclude that administration of the drug is feasible when they were only assessing survival).

### *Discussion*

Our results show that spin is prevalent in the abstracts of oncology clinical trials that measure OS and a surrogate endpoint. The conclusion sections of abstracts were most prone to contain spin. OS was more often a secondary endpoint. As a secondary endpoint, OS was statistically significant only twice; therefore, authors frequently concluded a treatment was effective based on significant surrogate endpoint data alone. And while it is not spin to discuss statistically significant surrogate endpoints when they are the primary endpoint of a trial, we considered it to be spin to ignore nonsignificant OS data and conclude that a treatment is definitively effective based on a surrogate endpoint alone. These results are in line with previous investigations of spin both in oncology[189,249,257] and the overall medical literature[137,242,258], indicating that spin continues to affect the accurate interpretation of trial results by physicians. The implication of this finding is that misrepresented, distorted research findings are being purported as true to oncologists.

Misrepresented or distorted research findings affect oncologist beliefs about drug efficacy. The SPIIN randomized trial demonstrated that oncologists believe drugs are more effective if the clinical trial abstract has spin in the conclusions.[52] Furthermore, oncologists are also more likely to read the full text of a clinical trial that has spin in the abstract. The reading habits of oncologists indicate that the effects of spin may be compounded since investigations of spin in the full text of trials have demonstrated that spin occurs at a similar rate.[137] It is known that internists often read only study abstracts, either to quickly learn or to screen out uninteresting results.[246] If these findings hold true for oncologists, trial authors may be incentivized to spin nonsignificant results, since spin leads to more interest among readers, and may improve the chances of favorable peer reviews and publication.

The tendency for trial authors to emphasize statistically significant surrogate endpoints when OS is nonsignificant is not surprising. Surrogate endpoints are increasingly important to the field of oncology and often are the basis for new drugs approvals under the FDA Accelerated Approval pathway.[3] Clinical trial authors may believe that the intervention drug is truly effective, even without OS data. Optimism bias toward new oncology interventions has been described previously.[257] However, when OS data is mature, available, and nonsignificant it may be difficult for authors to conclude that their drug is beneficial. Even if OS is a secondary endpoint, concluding that a treatment is beneficial may be difficult when only significant surrogate endpoint data are available at the time.

When authors deemphasize available OS data, there may be consequences for patients. Patients receiving adjuvant sunitinib may, like the oncologist-authors and the FDA reviewers, believe the drug is more effective than it truly is. Surrogate endpoints are useful when they predict OS early and accurately, but surrogate endpoints are incapable of completely replacing OS. Caution may be warranted in a trial that has statistically significant surrogate endpoint data and nonsignificant OS data, regardless of which endpoint is primary or secondary, since OS is what the surrogate endpoint is trying to predict.

To conclude, this investigation of spin in the abstracts of oncology clinical trials measuring OS and a surrogate endpoint shows that spin is common. Further, as a secondary endpoint, OS was statistically significant twice, raising questions about trial design and the utility of OS as a secondary endpoint. Nevertheless, authors frequently conclude a treatment is effective based on only statistically significant surrogate endpoint data. Spin was most common in the conclusion sections of abstracts, where authors interpret their results. The consequences of spin may include confusion about the true efficacy of a drug for patients and the dissemination of distorted conclusions to oncologists.

This study is limited by the 1-year cross section that was chosen for analysis. It is possible that our results do not reflect the reporting of oncology trials outside the chosen time frame, including clinical trials that were published in 2018 and later. Readers should account for this limitation when interpreting our study results.

# CHAPTER VII

# A COMPARISON OF MATCHED INTERIM ANALYSIS PUBLICATIONS AND FINAL ANALYSIS PUBLICATIONS IN ONCOLOGY CLINICAL TRIALS

***This work was previously published in the* Annals of Oncology *with the following citation:***

Wayant C, Vassar M. A comparison of matched interim analysis publications and final analysis publications in oncology clinical trials. Ann Oncol. Published online October 11, 2018. doi:10.1093/annonc/mdy447

## *Introduction*

In medical research, hype is the early excitement surrounding promising interventions, despite a lack of substantial supporting evidence.[259] Hype often opposes reason, but it is common, such as when cancer drugs are heralded as “game changers” despite having been evaluated only in animals and not having received Food and Drug Administration (FDA) approval. Journalists often perpetuate such hype, though physicians also may be responsible.[45] Cancer researchers perpetuate hype by casting unfavorable (nonsignificant) results in a favorable light.[185,249] This dangerous practice leads physicians to overstate a drug’s efficacy.[52]

We argue that hype affects the oncology community in particular owing to the prevalence of surrogate endpoints. The popularity of surrogate endpoints among oncology trial sponsors and investigators has been met with caution from others in the medical community.[260–262] This skepticism exists because surrogate endpoints often fail to predict the clinical endpoints that are most important to patients: OS and quality of life.[61,250,263] Nevertheless, most new drugs that receive accelerated approval from the FDA are analyzed using surrogate endpoints,[61] and the market price for drugs approved based on surrogate endpoints does not differ from the price of drugs approved based on OS.[264] When drugs granted accelerated approval report the required follow-up data, it is frequently for another surrogate endpoint, which can generate hype.[3] The current process for drug approvals means that formal assessments of OS may be delayed until after the drug has been widely used in patient care.

One increasingly popular surrogate endpoint for OS is PFS. PFS is a composite endpoint that combines assessments of tumor progression and death from any cause.[255] In oncology studies, the strength of association between PFS and OS varies and depends on tumor type, tumor stage, and drug intervention.[265,266] A recent study showed that effect sizes are significantly larger for PFS than for OS.[253] PFS requires less time and fewer patients to achieve statistical power. Thus, although oncology trials often report both PFS and OS, they may publish the mature PFS data apart from the mature OS data. Publishing interim trials with only mature PFS data apart from the confirmatory analyses of OS may promote hype.

In this investigation, we analyzed the hype generated by oncology clinical trials that published interim analyses. We investigated whether significant differences existed between interim and final analyses, with respect to the Altmetric score and journal prominence. We restricted our sample to trials that assessed both PFS and OS, because a recent analysis of interim results excluded these trials.[128]

### *Methods*

First, we searched PubMed, including Medline, on January 18, 2018, using the following search strategy: (((interim) OR immature) OR not mature) AND (overall survival AND progression free survival) AND Clinical Trial[ptyp] AND ("2005/01/01"[PDat] : "2015/12/31"[PDat]). We placed no restriction on included journals. All records were gathered in a PubMed collection.

Next, we exported this collection of 393 records to Rayyan.[256] We excluded any record that was not a randomized clinical trial, trial protocols, trials that did not assess both PFS and OS, and any record not available in English. To be included, a randomized clinical trial must have reported mature PFS data and denoted their OS data as immature. We further included any trials in which OS not denoted as immature but in which the prespecified number of deaths had not occurred. If we identified a trial reporting final or updated OS results, we attempted to identify the interim analysis with PFS data.

To match interim analyses with their corresponding final analyses, we used a combination of a PubMed search using the PICO (population, intervention, comparator, outcome) format, review of clinical trial registries, and emails to corresponding authors. If an author did not respond to our email, we sent 2 additional emails at 1-week intervals.

One of us (CW) independently extracted the following data from each interim and final analysis: title, year of publication, journal, intervention drug, comparator drug(s), whether a PFS benefit was demonstrated, median survival times and hazard ratio, cancer, cancer setting, trial funding source, trial design (e.g., superiority or noninferiority), whether an interim analysis was prespecified, and number of required PFS and OS events to achieve statistical power. All data were extracted via a piloted and validated Google Form. We retrieved the Altmetric score for each interim and final analysis using the "Altmetric it" bookmarklet, which identified the Altmetric score for each trial from PubMed.

To compare the interim and final analyses, we identified whether a PFS or OS benefit was demonstrated. Authors reported hazard ratios (HR) with confidence intervals most often. We calculated the ratio of hazard ratios (rHR) between PFS and OS (HR-PFS / HR-OS) to determine whether the HR effect size favored PFS or OS. This method was derived from a recent analysis of PFS and OS HR effect sizes.[253] As needed, the direction of effect was stabilized so that an HR less than 1.0 favored the intervention. This convention means that an rHR of less than 1.0 indicates a larger effect size for PFS compared to OS.

We further asked whether the publication of an interim analysis, apart from its final OS analysis, was justified based on the strength of the correlation between PFS and OS. To answer that question, we searched the literature to determine the strength of the correlation between the two endpoints for the specific tumor, tumor setting, and drug class. We began by referencing an SR of the correlation between surrogate endpoints and OS in oncology.[61] For any tumors, tumor settings, or drug classes that were not included, we searched PubMed (Medline) using the PICO format. We considered a strength of correlation of $r \leq 0.7$ to be low, $0.7 < r < 0.85$ to be medium, and $r \geq 0.85$ to be high, based on Prasad et al.'s SR, which adapted the Institute of Quality and Efficiency in Health Care's convention for trial-level correlation.[61,267]

We used the Wilcoxon signed-rank test for matched pairs to compare differences in continuous variables and Fisher's exact test to compare differences in categorical variables. For our rHR analysis, visual inspection of the histogram and results from the Shapiro-Wilk test of normality indicated that the data were not normally distributed. Therefore, we report a median rHR and use this median to determine the median-effect size difference between PFS and OS. We used Stata 13.1 for all statistical analyses.

### *Results*

Our search of PubMed retrieved 393 records. Figure 8 itemizes the 360 excluded records. We identified 27 interim analyses with mature PFS data, and we found an additional 6 via final analyses with mature OS data from our search. Of the 33 interim analyses with mature PFS data identified, only 25 could be paired with a final analysis. We excluded 2 of these matched pairs, because their final analyses were either in an abstract or in-press in a journal without an impact factor. The eight unmatched interim analyses have not yet published mature OS data. Therefore, 23 matched pairs were included. Unless otherwise specified, our results are for the 23 interim analyses with a final analysis pairing.

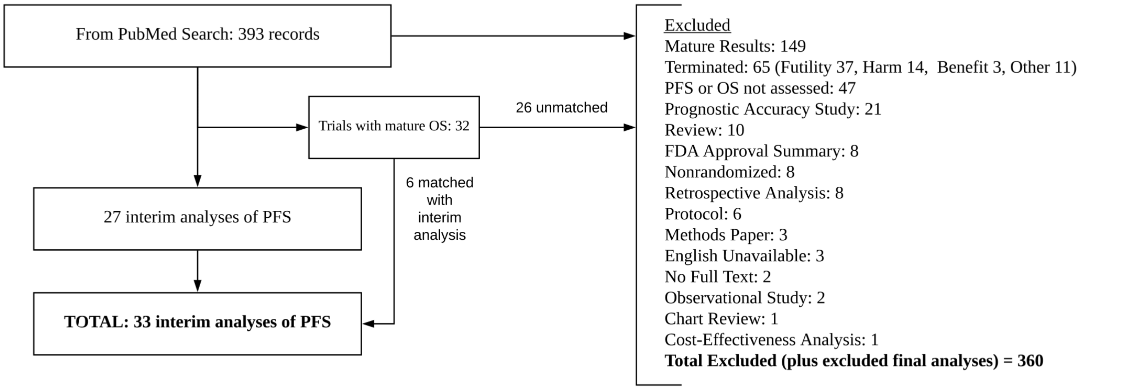
From PubMed Search: 393 records
Trials with mature OS: 32
26 unmatched
6 matched with interim analysis
27 interim analyses of PFS
TOTAL: 33 interim analyses of PFS
Excluded
Mature Results: 149
Terminated: 65 (Futility 37, Harm 14, Benefit 3, Other 11)
PFS or OS not assessed: 47
Prognostic Accuracy Study: 21
Review: 10
FDA Approval Summary: 8
Nonrandomized: 8
Retrospective Analysis: 8
Protocol: 6
Methods Paper: 3
English Unavailable: 3
No Full Text: 2
Observational Study: 2
Chart Review: 1
Cost-Effectiveness Analysis: 1
Total Excluded (plus excluded final analyses) = 360

*Figure 8. Flow diagram of included and excluded studies, including how studies were matched after initial search.*

All interim analyses ($n$ = 33) were prespecified and conducted in accordance with that trial's protocol. A statistically significant PFS benefit occurred in 93.9% (31/33) of interim analyses. At long-term follow up, the PFS effect size decreased in 8 trials (2 became nonsignificant), increased in 5 trials, and remained the same in 1. Eleven trials did not report updated PFS data. A statistically significant benefit in OS occurred only 8 times, although 12 trials allowed crossover and 2 administered additional therapies to patients after progression. In the 2 interim analyses with a statistically nonsignificant PFS benefit, a statistically significant OS benefit occurred in 1 of them. That trial compared concurrent and sequential alternating gefitinib in previously untreated metastatic non-small cell lung cancer.[268] Across all studies (*n* = 33), to achieve statistical power, the median number of required patient events was 282.50 (interquartile range [IQR] 191.50–380.25) for interim PFS analyses and 385 (IQR 245–492) for final OS analyses.

Among matched pairs (*n* = 23), interim analyses were published in more prominent journals compared to final analyses (Table 6). Specifically, interim analyses were more likely to be published in the top 5 general medicine journals (e.g., *New England Journal of Medicine*, *The Lancet*) but not more likely to be published in the top 5 oncology journals (e.g., *Journal of Clinical Oncology*, *The Lancet: Oncology*). Interim analyses were published in journals with an impact factor of ≥ 20 more often, but this difference was not statistically significant. The median impact factor was 44 (IQR 24–72) for journals publishing interim analyses versus 24 (IQR 11–34) for journals publishing final analyses. Only 2 interim analyses were published in journals with an impact factor less than 10, compared to 6 final analyses; 1 final analysis was published as an abstract, and 1 was an in-press manuscript. The impact factor increased once and remained the same 3 times from interim to final publication.

Interim analyses also had higher Altmetric scores than final analyses. The median Altmetric score was 28 (IQR 13.25-82.25) for interim analyses versus 18 (5-46) for final analyses (*p* = .002). Of the 2 final analyses that had no Altmetric scores or impact factors, 1 was an abstract and 1 was an in-press manuscript in *ESMO: Open*. The Altmetric score increased 3 times and remained the same 1 time from interim to final analyses.

We were able to compare PFS and OS effect sizes in 24 trials. The PFS effect size was larger than the OS effect in 21 of 24 (87.5%) trials. When comparing the interim analyses with mature PFS data and the final analyses with mature OS data, the median rHR was 0.69 (0.51–0.86), corresponding to a median 31% larger effect size for PFS compared to OS (Table 7). The rHR was the same in immunotherapy trials (n = 8) where traditional

PFS is not yet a validated surrogate endpoint [0.69 (IQR 0.66-0.82)]. The effect size for PFS was larger than OS in all matched pairs of immunotherapy trials.

| **Characteristics** | | **n (%)** |
|---|---|---|
| *Reason for Interim Analysis* | | 33 (100) |
| | Impact Factor ≥20 | 29 (87.9) |
| | Top 5 General Medical* | 18 (54.5) |
| | Top 5 Oncology** | 11 (33.3) |
| *Altmetric Score* [Median (IQR)] | | 50 (10 - 129) |
| *PFS Benefit* | | |
| | Significant | 31 (93.9) |
| | Nonsignificant | 2 (6.1) |
| *Funding* | | |
| | Industry Alone | 27 (81.8) |
| | Industry Partly | 2 (6.1) |
| | Government | 1 (3.0) |
| | Non-Profit | 1 (3.0) |
| | None Listed | 2 (6.1) |
| *Design* | | |
| | Superiority | 32 (97.0) |
| | Noninferiority | 1 (3.0) |
| | Phase II | 3 (9.1) |
| | Phase II/III | 1 (3.0) |
| | Phase III | 29 (87.9) |
| *Control Group* | | |
| | Active | 13 (39.4) |
| | Placebo | 6 (18.2) |
| | Placebo Plus Active | 13 (39.4) |
| | Surgery | 1 (3.0) |
| *Number of PFS Events* [Median (IQR)] | | 287 (193 - 384) |
| Strength of Correlation Between PFS & OS | | |
| | High (r ≥ 0.85) | 3 (9.1) |
| | Medium 0.7 < r < 0.85) | 5 (15.2) |
| | Low r ≤ 0.7 | 22 (66.7) |
| | Unknown | 3 (9.1) |

*Table 6. Characteristics of all interim analyses of progression-free survival (n = 33).*

| Characteristics | Publication | | |
|---|---|---|---|
| | Interim (PFS) | Final (OS) | Difference (p-value) |
| *Journal Prominence** | | | |
| Impact Factor ≥20, n (%) | 21 (91.3) | 15 (65.2) | 0.07 |
| Top 5 General Medical, n (%) | 14 (60.9)) | 3 (13.0) | <0.01 |
| Top 5 Oncology, n (%) | 8 (34.7) | 13 (56.5)) | 0.26 |
| Both Interim & Final in Impact ≥20 Journal | | 13 (56.5) | |
| *Altmetric Score* [Median (IQR)]* | 28 (13.25 - 82.25) | 18 (5 - 46) | <0.01 |
| *Interim PFS (+) [n=15], Final OS (-) [n = 14]** | | | |
| Impact Factor ≥20, n (%) | 12 (80.0) | 5 (35.7) | <0.03 |
| Altmetric Score [Median (IQR)] | 16.5 (7.25 - 43.25) | 7 (1.5 - 17.25) | <0.05 |
| *Interim PFS (+), Final OS (+) [n = 8]* | | | |
| Impact Factor ≥20, n (%) | 8 (100) | 8 (100) | n/a |
| Altmetric Score [Median (IQR)] | 137 (82 - 161) | 42 (33 - 84) | <0.05 |
| *Interim PFS (-), Final OS (-) [n = 1]* | | | |
| Journal | The Lancet: Oncology | Journal of Clinical Oncology | n/a |
| Altmetric Score | 77 | 60 | n/a |
| *Interim PFS (-), Final OS (+) [n = 1]* | | | |
| Journal | Annals of Oncology | ESMO Open | n/a |
| Altmetric Score | 6 | n/a | n/a |

*Table 7. Characteristics of matched pairs of interim PFS and final OS analyses (n = 23). *Two were excluded from overall matched pair analysis due to publication as an abstract or for being in-press in a journal without an Impact Factor. The in-press matched pair is included in the stratified analysis but its Impact Factor is shown as "n/a".*

In 19 of the 25 total matched pairs, there was a low ($r \leq 0.7$) or unknown strength of correlation between PFS and OS (Table 8). In 3 cases, the strength of correlation was medium ($0.7 < r < 0.85$). In 3 other cases, the strength of correlation was high ($r \geq 0.85$). For the 8 interim analyses that had yet to report final OS data, the strength of correlation between PFS and OS was low in 5 and medium in 3. All 8 unmatched interim analyses were published in *New England Journal of Medicine* (*n* = 3), *The Lancet: Oncology* (*n* = 2), *Journal of Clinical Oncology* (*n* = 2), or *The Lancet* (*n* = 1).

Overall, *New England Journal of Medicine* published the most interim analyses of PFS (*n* = 9), all of which were statistically significant. Only 2 nonsignificant interim analyses of PFS were published — 1 in *The Lancet: Oncology* and 1 in *Annals of Oncology.* For final analyses of OS, *The Lancet: Oncology* published the most (*n* = 7), of which 5 were statistically significant.

|  | Title | Intervention | PFS Hazard Ratio | OS Hazard Ratio | Ratio of Hazard Ratios | Type of Cancer | Strength of Correlation |
|---|---|---|---|---|---|---|---|
| **PFS (+)<br>OS (-)<br>n = 16** | Yardley, 2013 | Everolimus | 0.38 (0.31 - 0.48) | 0.89 (0.73 - 1.10) | 0.43 | Breast | Low |
|  | Schmittel, 2006 | Irinotecan | no HR | 0.75 (0.54 - 1.03) | n/a | Small-cell lung | Medium |
|  | Brufsky, 2012 | Bevacizumab | 0.49 (0.33 - 0.74) | 1.01 (0.85 - 1.22) | 0.49 | Breast | Low |
|  | Pujade-Lauraine, 2010 | Pegylated liposomal doxorubicin | 0.82 (0.72 - 0.94) | 0.99 (0.85 - 1.16) | 0.83 | Ovarian | Unknown |
|  | Aghajanian, 2012 | Bevacizumab | 0.48 (0.39 - 0.61) | 0.95 (0.77 - 1.18) | 0.51 | Ovarian, Peritoneal, or Fallopian tube | Unknown |
|  | Markman, 2003 | Paclitaxel | 2.3 (1.08 - 4.94) | 0.91 (0.70 to 1.14) | 2.54 | Ovarian | Low |
|  | Escudier, 2007 | Bevacizumab | 0.63 (0.52 - 0.75) | 0.91 (0.76 - 1.10) | 0.69 | Renal | Low |
|  | Pavel, 2011 | Everolimus | 0.77 (0.59 - 1) | 1.17 (0.92 - 1.49) | 0.66 | Neuroendocrine | Low |
|  | Nordlinger, 2008 | FOLFOX4 | 0.73 (0.55 - 0.97) | 0.88 (0.68 - 1.14) | 0.83 | Colorectal | High |
|  | Bolla, 2005 | Post-op irradiation | 0.49 (0.41 - 0.59) | 1.18 (0.91 - 1.53) | 0.42 | Prostate | Unknown |
|  | Motzer, 2008 | Everolimus | 0.3 (0.22 - 0.40) | 0.87 (0.65 - 1.15) | 0.34 | Renal | Low |
|  | San-Miguel, 2014 | Panobinostat | 0.63 (0.52 - 0.76) | 0.94 (0.78 - 1.14) | 0.67 | Multiple myeloma | Medium |
|  | Ribas, 2015 | Pembrolizumab [2mg/kg arm] | 0.57 (0.45 - 0.73) | 0.86 (0.67-1.10) | 0.66 | Melanoma | Low |
|  |  | Pembrolizumab [10mg/kg arm] | 0.50 (0.39 - 0.64) | 0.74 (0.57-0.96) | 0.68 | Melanoma |  |
|  | Perren, 2011 | Bevacizumab | 0.81 (0.70 - 0.94) | 0·99 (0.85 - 1.14) | 0.82 | Ovarian | Unknown |
|  | Ledermann, 2012 | Olaparib | 0.35 (0.25 - 0.49) | 0.73 (0.55 - 0.96) | 0.48 | Ovarian | Unknown |
| **PFS (+)<br>OS (+)<br>n = 8** | Ryan, 2013 | Abiraterone | 0.53 (0.45 - 0.62) | 0.81 (0.70-0.93) | 0.65 | Prostate | Unknown |
|  | Choueiri, 2015 | Cabozantinib | 0.58 (0.45 - 0.75) | 0.66 (0.53 - 0.83) | 0.88 | Renal | Low |
|  | Stewart, 2015 | Carfilzomib | 0.69 (0.57 - 0.83) | 0.79 (0.67 - 0.95) | 0.87 | Multiple myeloma | Medium |
|  | Long, 2014 | Trametinib | 0.75 (0.57 - 0.99) | 0.71 (0.55 - 0.92) | 1.06 | Melanoma | High |
|  | Larkin, 2014 | Cobimetinib | 0.51 (0.39 - 0.68) | 0.7 (0.55 - 0.90) | 0.73 | Melanoma | High |
|  | Baselga, 2012 | Pertuzumab | 0.62 (0.51 - 0.75) | 0.66 (0.52 - 0.84) | 0.94 | Breast | Low |
|  | Escudier, 2007 | Sorafenib | 0.44 (0.35 - 0.55) | 0.88 (0.74 - 1.04) | 0.50 | Renal | Low |
|  | Krop, 2014 | Trastuzumab | 0.53 (0.42 - 0.66) | 0.68 (0.54 - 0.85) | 0.78 | Breast | Low |
| **PFS (-)<br>OS (-)** | Weber, 2015 | Nivolumab | 0.82 (0.32 - 2.05) | 0.95 (0.73 - 1.24) | 0.86 | Melanoma | Low |
| **PFS (-)<br>OS (+)** | Sugawara, 2015 | Gefitinib | 0.71 (0.42 - 1.20) | 0.58 (0.34 - 0.97) | 1.22 | Non-small cell lung | Low |

*Table 8. Comparison of Hazard Ratios for PFS and OS Among Matched Pairs (n=25). A hazard ratio (HR) less than one favors the intervention. The ratio of hazard ratios (rHR) is equal to (HR-PFS / HR-OS). A rHR less than one indicates a larger effect size for PFS.*

## *Discussion*

Our results demonstrate that interim analyses with mature PFS data generate hype in oncology. Compared to final analyses, interim analyses are more likely to be published in top-5 general medical journals and more likely to have higher Altmetric scores. Two factors help to explain these differences. PFS was more likely to be statistically significant and have a larger effect size. Additionally, two recent investigations show that interim analyses in oncology are associated with exaggerated effect sizes and that PFS effect sizes are larger than OS effect sizes.[253,269] Another investigation found that for FDA-approved drugs, the post-approval trials frequently have smaller treatment effects compared to their matched pre-approval trials.[270] Lastly, Woloshin, et al found that one-fifth of final analyses fail to agree with the conclusions of the interim analyses.[128] This last study suggests that no significant difference exists between interim and final publications regarding journal prominence and Altmetric score. Our results show the opposite.

We believe that hype in the oncology literature is easier to generate than in the overall medical literature for several reasons. First, most current FDA approvals are for oncology drugs.[271] Also, social and governmental pressure creates a sense of urgency among the oncology community to make cancer therapies available to patients.[46] Moreover, a large treatment effect often causes excitement, especially given the knowledge that recently approved cancer drugs improve OS by a median of only 2 months.[116] Nevertheless, our results should not discourage the analysis of surrogate endpoints. Instead, they should encourage the proper presentation of surrogate endpoint results.

The publication of interim analyses with only mature PFS data apart from confirmatory OS analyses must be cautioned. We have demonstrated that the PFS effect sizes are frequently exaggerated, compared to the final OS analyses, and that OS fails to confirm a significant PFS benefit in most cases. On the contrary, recent investigations of the correlation between PFS and OS in PD-1 inhibitor immunotherapy trials showed the opposite – that PFS effect sizes were smaller compared to OS effect sizes[272]. Moreover, in two cases — one investigating bevacizumab in ovarian cancer and one investigating perioperative FOLFOX4 in metastatic colorectal cancer — the significant PFS effect from the interim analyses became nonsignificant at follow-up[273,274]. The clinical relevancy of inhibited tumor growth, which likely contributes to the exaggerated PFS effect size, is relevant to patients only if the OS benefit follows. Our data show that the mature PFS data in interim analyses were more often statistically significant than subsequent OS data. Crossover was allowed after disease progression in 12 of the included trials and 2 trials administered additional interventions — which may mask true OS benefits. Whether crossover after disease progression is an acceptable feature of trial design has been debated recently[49,123,124].

We further demonstrated that the incidence of publishing interim analyses separately from their final confirmatory analyses has steadily increased. Numerous manuscripts have called for caution when interpreting surrogate endpoint data.[61,250,252,260–262] Our results support this cautionary call: only 3 of the 33 included interim analyses showed that PFS strongly correlated with OS. For any cancer, cancer setting, or drug intervention without a validated correlation of PFS and OS, the interim publication of PFS data without accompanying OS data is not just unreliable, but likely to generate hype. We, therefore, recommend caution when reading interim analyses with only PFS data, since the effect size and clinical benefit may not be corroborated by future OS data.

Analyses of surrogate endpoints likely affects clinical decision making. A recent review of FDA drugs that received accelerated approval in the last 25 years showed that all were approved based on a surrogate endpoint and that most received regular approval based on another surrogate endpoint.[3] The strength of correlation between surrogate endpoints and OS could be high for many of these drugs, but it is more likely that the OS effect size is small or null. To be certain, adding even 2 months to a patient's life cannot be discounted and is incredibly important, but patients often expect much more when choosing between treatment options.[275] One must question how the perceived survival benefit demonstrated by surrogate endpoints affects physician and patient expectations around treatment decisions.

The limitation of our study is that our results are not generalizable to all surrogate endpoints, because they may correlate with OS differently and may be published apart from OS analyses more or less frequently. Further, the magnitude of difference in Altmetric scores between interim and final analyses may be biased, because the final analyses were published more recently. However, only 6 final analyses were published in 2017 or later, and 1 was in-press at the time of analysis and thus excluded. Of the 5 that were included, 3 had Altmetric scores well above the median (69, 60, and 43). We are therefore confident that the bias due to time of publication is minimal.

To conclude, we do not discourage the use of surrogate endpoints in oncology. They are valuable and serve a useful purpose. We do, however, encourage the proper presentation of surrogate endpoint results in oncology. PFS effect sizes are frequently larger than OS effect sizes, and PFS is infrequently validated as a surrogate endpoint for OS. We recommend caution when encountering an oncology trial with only immature OS data, because we have demonstrated that such interim analyses may generate unsupported and inappropriate hype. When these interim analyses are published, the journals should provide timely links to the final publication, even if it is published in a different journal.

# CHAPTER VIII

# EVALUATION OF SELECTIVE OUTCOME REPORTING BIAS IN EFFICACY ENDPOINTS IN PRINT AND TELEVISION ADVERTISEMENTS FOR ONCOLOGY DRUGS

***This work was previously published in the* Journal of General Internal Medicine *with the following citation:***

Wayant C, Aran G, Johnson BS, Vassar M. Evaluation of Selective Outcome Reporting Bias in Efficacy Endpoints in Print and Television Advertisements for Oncology Drugs. J Gen Intern Med. 2020;35(10):2853-2857.

## *Introduction*

Industry-sponsored television and print advertisements targeted to consumers and health care providers (HCPs) compose a multibillion dollar industry in the United States.[276] Consequently, the benefits and harms of these advertisements have been strongly debated, with much of the discussion focusing on consumers.[277,278] Advocates of direct-to-consumer advertisements argue that they function as public service announcements that empower patients with information, lead to doctor-patient conversations, and facilitate the initiation of treatment.[279–281] Opponents argue that direct-to-consumer advertisements may mislead patients,[282,283] exaggerate potential drug benefits,[284,285] omit quality of life,[286] and increase health care spending.[279,280] In cancer medicine, drug advertisements have been the subject of particularly intense debate,[286–288] especially given the often high toxicity[289] and cost[290] associated with new cancer medications. The controversial nature of oncology drug advertisements, paired with their prevalence in the lives of HCPs and consumers, raises the critical question of whether the clinical data in oncology drug advertisements are transparent, straightforward, and unbiased.

One threat to the accurate presentation of clinical data is selective outcome reporting bias, which occurs when published study endpoints do not match those prespecified in a trial registry or protocol.[291] Trial endpoints may be added, removed, or reordered for several reasons. Some of these reasons, such as poor study accrual,[292] are ethical and understandable. However, in other cases, selectively reporting endpoints can be dangerous and may affect perceptions of drug efficacy through the omission or demotion of statistically nonsignificant results. A recent analysis of hematology clinical trials found that endpoints were often selectively reported to highlight statistically significant results,[36] and a Cochrane SR found that selective outcome reporting bias in clinical trials affected the conclusions of a "substantial proportion of Cochrane reviews."[293] To avoid misleading readers, authors of medical research studies should accurately report data for all endpoints prespecified in their protocol, regardless of statistical significance.

While much is known about the selective reporting of trial endpoints between protocols and published reports, little, if anything, is known about the selective reporting of trial endpoints between published reports and drug advertisements. Because advertisements represent a snapshot of a drug's evidence profile, they may be slanted toward selective reporting of endpoints previously analyzed in published trials. The primary objective of the current study was to investigate the rates of selective outcome reporting bias of efficacy endpoints at two junctures: in published cancer clinical trials and in television and print advertisements for anticancer medications. The rationale for this investigation was that selective outcome reporting bias has been shown to be a consistent issue in the biomedical literature,[36,136,293,294] and print or television advertisements may unintentionally inflate perceptions of the benefits of oncology drugs.

### *Methods*

Consistent with a recent investigation of health care advertisements,[295] we used the AdPharm database to identify oncology drug advertisements uploaded within an 18-month span between March 1, 2017, and September 1, 2018. AdPharm is an online database that is updated daily with advertisements for health care or pharmaceutical products. Each entry in AdPharm contains basic information about the advertisement, including the target audience or country of origin. AdPharm does not track or list the number of viewers of an advertisement. Advertisements were included in the study if they were for an anticancer drug and if they included quantitative data, were in English, and were marketed to consumers or HCPs.

After screening all advertisements, CW and GA extracted data in a duplicate and masked fashion. The following items were extracted from print and television advertisements: market audience, air or print date, efficacy endpoints, data for efficacy endpoints, design features of the clinical trial that generated the data, any citation for a published trial, and, in the case of a consumer-directed advertisement, any mention of speaking with an HCP.

To compare advertisement endpoints with journal-published endpoints, we used the citations in the advertisements or a PubMed search to identify a matching trial. We used keywords and Boolean operators to search for and identify matching trials, if no citation was included. Trials were matched on the basis of intervention, co-intervention, control, sample size, and cancer type. After identifying matched trials, we extracted the efficacy endpoints reported, data for those endpoints, and the date of article publication. Our analysis of selective reporting bias between published articles and advertisements consisted of determining which endpoints were included in the published paper and which were included in the advertisements. When an endpoint was excluded from the advertisement, we then determined whether or not that endpoint was statistically significant using the published statistics (e.g., confidence intervals or alpha level). Similarly, we investigated selective outcome reporting between the retrieved published papers and their trial registrations. We chose to use trial registrations, rather than protocols, because trial registrations are time-stamped and show a history of changes, which supports an accurate analysis of any endpoint changes or updates.

This is a novel study of selective outcome reporting in drug advertisements. As such, there is no effect size on which to base a power calculation. Therefore, we provide a range of included studies required for sufficient power using standard effect size measurements (Cohen's d = 0.2, 0.5, 0.8). These effect size measurements were converted to odds ratios for our power calculation, based on the paper by Chen, et al[296]. We prespecified a type I error rate of 0.5 and type II error rate of 0.2. The range of included advertisements required ranged from 485 (odds ratio = 1.68, Cohen's d = 0.2) to 89 (odds ratio = 3.47, Cohen's d = 0.5) to 45 (odds ratio = 6.71, Cohen's d = 0.8). We used gpower 3.1 for all power calculations.

We used Stata 15.1 for all analyses except E-values, for which we relied on the formula described by VanderWeele and Ding.[297] E-values were used to assess the degree of unmeasured confounding in our analyses. For the two primary endpoints of selective outcome reporting bias of efficacy endpoints in published papers and in advertisements, we calculated unadjusted risk ratios (uRR) and 95% confidence intervals (CIs) to compare the rates of advertising significant and nonsignificant endpoints. We analyzed all advertisements together, as well as consumer- and physician-directed advertisements separately. In all analyses of selective outcome reporting bias, we excluded endpoints from single-arm trials, immature OS data, and endpoints that could not be located in the published paper. We define "immature" data as data that have not accrued the prespecified number of patient events to achieve study power.

***Results***

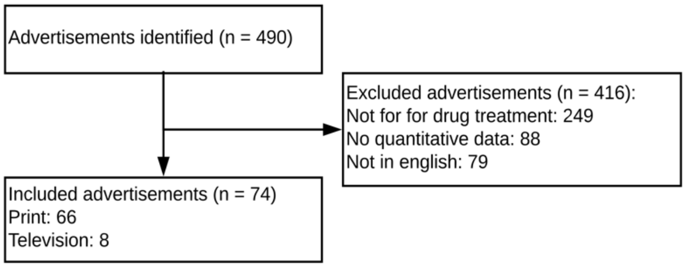
Advertisements identified (n = 490)
Excluded advertisements (n = 416):
Not for for drug treatment: 249
No quantitative data: 88
Not in english: 79
Included advertisements (n = 74)
Print: 66
Television: 8

*Figure 9. Flow diagram of included studies.*

We identified 490 advertisements in total, of which 74 were included in initially (Figure 9). Advertisements were excluded for not describing a drug treatment (n = 249), not including quantitative data (n = 88), and not being in English (n = 79). The vast majority of print advertisements (n = 66) were in clinical magazines and designed to target HCPs (n = 55, 83.3%). Print advertisements pertained to 34 unique drugs designed to treat 21 unique malignancies. The drugs that were the most commonly advertised in print were pembrolizumab (n = 8), palbociclib (n = 6), and ribociclib (n = 5). All television advertisements (n = 8) were directed to consumers and were related to four unique drugs and two unique malignancies. Palbociclib was the most commonly television-advertised drug (n = 3), followed by pembrolizumab (n = 2), nivolumab (n = 2), and abemaciclib (n = 1). The only malignancies represented were non-small-cell lung and breast cancers (both n = 4).

### *Registration to Publication*

Forty-eight clinical trials were identified that supported the 74 included advertisements. All 48 trials reported a trial registration number. Seven trials were registered after the start of subject enrollment, although one trial began in 1999 before ClinicalTrials.gov registration. Besides the six trials that were registered after they began (excluding the trial that began in 1999), an additional six studies deviated from the registered primary endpoints in ways that may have affected the integrity of the trial. For all six, primary endpoints were added to the registry after the start of the study. In one study, an endpoint was demoted from primary to secondary in the published report. With regard to registered secondary endpoints, 16 trials deviated from the registry, with 13 adding secondary endpoints during the trial period. One study promoted a registered secondary endpoint to a primary endpoint in the publication, one removed a secondary endpoint from its registry, and one did not list or report a registered secondary endpoint in the paper. Overall, 41/48 (85.4%) trials were registered prior to study enrollment and 41/48 (85.4%) did not deviate from the registered primary endpoints.

### *Publication to Advertisement*

After excluding advertisements supported by single arm trials (n = 8), we next compared the efficacy endpoints cited in the 66 remaining advertisements to the 40 remaining clinical trials supporting them. Of the 539 endpoints eligible for inclusion in advertisements, we excluded 175 endpoints for being from single-arm trials (n = 100), for including immature time to event data (n = 51), or for not including a statistical analysis in the published paper (n = 24). Five trials were cited for advertisements directed to consumers and physicians.

Across all included advertisements (n = 66), statistically significant endpoints were more likely to be reported than nonsignificant endpoints (uRR 1.26; 95% CI 1.14—1.40). Primary endpoints were reported 97.8% (92/94) of the time. Secondary endpoints were reported much less frequently (66/270, 24.4%). Overall, half (33/66, 50.0%) of advertisements included data for immature endpoints.

Among advertisements directed to HCPs (n = 47), if an endpoint was statistically significant, it was more likely to be reported in the advertisement (uRR 1.36; 95% CI 1.20–1.54). For consumer-directed advertisements, there was no significant difference (uRR 1.01; 95% CI, 0.85–1.21).

### *Discussion*

This study is a novel investigation of selective outcome reporting in drug advertisements marketed to consumers and health care providers. We found that statistically significant endpoints were more likely to be reported than nonsignficant endpoints. This finding was mostly driven by physician-directed advertisements, which were more prevalent and where the difference was also significant. Because previous studies investigating selective outcome reporting in drug advertisements do not exist, it is not possible to compare our results within the context of previous literature. In this study, we also evaluated selective outcome reporting between trial registrations and the published trial reports, which is the conventional manner for the investigation of selective outcome reporting [134,291,298]. There is ample evidence that industry-funded studies are more likely to report more favorable results in published papers[224,299,300]. Our results indicate that the degree of selective outcome reporting was higher between published trial reports and advertisements than between the trial registrations and their publications. These findings raise important questions about perceptions of drug efficacy. Moreover, many included endpoints were surrogate endpoints, which may  or may not correlate with improved survival in cancer patients[62] and are more likely to be statistically significant[204]. Some cancer trialists have argued that OS should be routinely collected and reported, owing to the importance that patients with cancer place on decreased mortality[53].

Our study found that advertisements were often aired or printed before final OS data were available, which may introduce uncertainty and may raise the risk of reporting false-positive results to the public [301]. Previous studies have found that only negligible correlations exist between surrogate outcomes and OS for many types of cancer [62]. Further, the results from surrogate outcomes—published as interim analyses before OS data are mature—often do not result in improvements in OS [204]. Thus, we believe that the surrogate outcomes reported in media advertisements have the potential to overstate the efficacy benefit that will eventually be found when OS data become available.

To our knowledge, the Food and Drug Administration (FDA) does not offer guidance on reporting surrogate endpoints and OS in oncology drug advertisements. Existing draft guidance for advertising efficacy endpoints focuses on the reporting of absolute or relative statistics.[302] This gap in FDA guidance may be relevant to patients if advertisements only report surrogate endpoints. A recent review found that there are no high-quality data supporting the idea that patients understand surrogate endpoints and their shortcomings.[58] The lack of guidance and patient misunderstanding may multiply issues with oncology drug advertisements. Namely, we have shown that nonsignificant endpoints and immature OS data are often excluded from oncology drug advertisements, resulting in a higher degree of significant surrogate endpoints, which patients may not fully understand.

One must weigh the benefits and harms of oncology drug advertisements as seen in this study. The advertisements that we assessed often excluded nonsignificant endpoints, yet drug advertisement proponents argue that advertisements, any selective reporting aside, initiate a patient-physician conversation.[279–281] Because of the paucity of research into the effects of selective outcome reporting in drug advertisements, we cannot address how the omission of nonsignificant endpoints affects patients' perceptions of drug efficacy. It is even more difficult to assess the effect of these advertisements on physicians, who in theory should be well versed in clinical endpoints and should have read the clinical trials associated with advertised drugs. However, we believe our study raises questions that could be answered using robust methodologies, following the example of other forms of bias [52].

Our study is limited because we were not able to assign an appropriate weight to each advertisement based on audience size. The television advertisements, which were all marketed to consumers, likely had a larger audience than the print ads, which were mostly marketed to HCPs and published in clinical journals. So, while most advertisements included in this study were marketed to HCPs, these advertisements were likely seen by fewer people. Moreover, it is difficult to determine whether selective outcome reporting in patient-directed advertisements (where it exists) has the same effect as in physician-directed advertisements. We may reasonably assume a higher degree of health literacy among physicians; therefore, selective reporting of endpoints in advertisements directed to physicians may not carry similar weight as for advertisements directed to patients. Last, the computed E-values for this study show that unobserved confounding may affect our results (i.e., that some factor other than the significance of endpoints may drive reporting). However, even if other factors contribute to the reporting of endpoints in advertisements, this finding does not change the fact that we identified a possibly biased drug efficacy portfolio in advertisements.

In conclusion, we found that oncology drug advertisements are more likely to include statistically significant endpoints than nonsignificant endpoints. This effect was most pronounced in advertisements marketed to HCPs. All advertisements relied mostly on surrogate endpoints and frequently omitted nonsignificant OS data. Immature OS data did not create a barrier to advertising a drug as effective to consumers and HCPs. We recommend that advertisements not be aired or printed without clear descriptions of patient-important endpoints, such as OS. Further, we recommend that the FDA critically review advertisements in the preapproval stage to ensure that patients and physicians are not misled (even unintentionally) regarding drug efficacy. We advocate for improved patient education of surrogate endpoints because available studies have shown that patients may conflate surrogate endpoints with clinically meaningful outcomes.[58] At minimum, since few, if any, oncology drugs aim to improve quality of life alone, we recommend a clear, prominent declaration of whether or not the drug has shown OS improvements. Future studies should be conducted to confirm our results, using a larger cohort of advertisements.

# CHAPTER IX

## METHODOLOGICAL QUALITY OF ONCOLOGY NONINFERIORITY CLINICAL TRIALS

***This work was previously published in* Critical Reviews in Oncology/Hematology *with the following citation:***

Wayant C, Ross A, Vassar M. Methodological quality of oncology noninferiority clinical trials. Crit Rev Oncol Hematol. 2020;149:102938.

---

### *Introduction*

Noninferiority trials compare a new treatment (NT) against an active control (AC) and aim to demonstrate that the NT is not worse than the AC, within a certain margin.[303] NTs that demonstrate noninferiority versus an AC usually exhibit a tradeoff: slightly less (but acceptable) therapeutic efficacy in exchange for increased safety or decreased cost. However, the design, reporting, and interpretation of noninferiority trials has been questioned,[304–307] and trials that are poorly designed, reported, and interpreted may represent a breach of ethical obligations to patients.[308] Another consequence may be spurious conclusions of noninferiority.

Basic design characteristics and the choice of noninferiority margin (i.e., the acceptable amount of efficacy lost) are some of the key issues in noninferiority trials. Basic design characteristics may include the choice of type I error rate (alpha) and the number of analyses run on the primary endpoint.

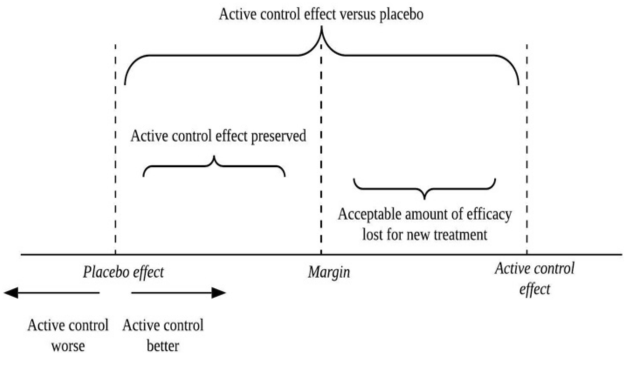
Active control effect versus placebo
Active control effect preserved
Acceptable amount of efficacy lost for new treatment
Placebo effect
Margin
Active control effect
Active control worse
Active control better

*Figure 10. Visual representation of the noninferiority margin as it relates to the expected effects of new treatments and active controls.*

Lenient alpha values or alpha values that do not align with the reported confidence intervals may compromise noninferiority conclusions.[304] Simultaneous intention-to-treat and per protocol analyses are encouraged in noninferiority trials[309,310] because reporting only intention to treat (the gold standard for superiority trials) may bias the results toward conclusions of noninferiority. Each of these basic design characteristics is as important to robust noninferiority conclusions as the choice of noninferiority margin. The margin represents the acceptable loss of efficacy for the NT compared to an AC that is more effective but may lack nonefficacy benefits of the NT (Figure 10). If the chosen margin of efficacy difference is too wide, then whatever effect the AC previously showed against placebo or another historical control may be lost. Even if the margin is appropriate, there may be concerns, as evidenced by the approval of lenvatinib for hepatocellular carcinoma.[311] Lenvatinib showed noninferiority to sorafenib, which had previously shown superiority to a placebo; however, it was later found that the clinical trial efficacy of sorafenib did not generalize to older patients and patients with worse performance status.[312] Thus, the supposed efficacy of sorafenib, which formed the basis for the chosen noninferiority margin, resulted in a conclusion about noninferiority that is not robust.

Building on recent investigations of noninferiority trials across biomedicine[304,313,314] in the context of an increasing number of new oncology drug approvals based on noninferiority trials, we aimed to determine the robustness of noninferiority trial design in cancer medicine. Our investigation had 3 components: (1) investigate basic design characteristics, (2) determine the percentage of AC effect preserved by the chosen margin, and (3) assess the overall quality of included trials using data gathered from components 1 and 2.

### *Methods*

Our protocol with detailed methods and search strategy is available via the Open Science Framework.[315] Briefly, we used a PubMed search to collect oncology noninferiority trials published in *New England Journal of Medicine*, *The Lancet*, *Journal of Clinical Oncology*, *Lancet Oncology*, *JAMA Oncology*, *Annals of Oncology*, *Cancer*, *European Journal of Cancer*, *British Journal of Cancer*, and *Journal of the National Cancer Institute* between January 1, 2012, and December 31, 2018. All retrieved studies from our search were exported to Rayyan,[256] a web-based platform used to screen studies for eligibility.

All articles were independently screened by 2 masked authors. Studies were included if they represented an oncology, noninferiority, phase 3 randomized controlled trial (RCT) of an antitumor or adjunct therapeutic (e.g., colony-stimulating factor) intervention. We excluded phase 2 studies, studies that only assessed the superiority of an intervention, studies that did not evaluate antitumor or adjunct therapeutic interventions, studies that used Bayesian methods, and studies that had a design other than an RCT.

Data extraction was performed independently by 2 masked authors using 2 Google Forms. The first form included items related to the basic characteristics and design of the noninferiority trial (see our protocol for the full list of items).[315] The second form was used to extract data from the noninferiority studies that cited previous studies of the AC versus a placebo or another control to justify the noninferiority margin. These previous results must have been used to justify the noninferiority margin, and they must have tested the same AC (including dose and administration procedures) against a placebo or another control for the same endpoint used in the noninferiority trial. The following items were extracted from the previous AC study into this second form: effect size, confidence interval, and *P*-value.

If multiple background studies of the AC versus placebo or other control were referenced in the noninferiority trial, we followed the algorithm devised by Tsui et al[314] to select 1 background study because of the inherent difficulties and limitations of attempting to generate a single effect size from multiple heterogeneous studies. The algorithm is a set of hierarchical criteria, in order of decreasing importance: (1) similarity of AC in the noninferiority trial and background study (e.g., dose, regimen); (2) placebo-controlled studies preferred over other control studies; (3) similarity of outcome between noninferiority trial and background study; (4) higher-order studies preferred over primary studies (e.g., meta-analysis over RCT); and (5) more recent study preferred.

For each noninferiority trial that cited a previous study with data that could be extracted into the second Google Form, we calculated the percentage of preserved effect (%PE), or the minimum effect of the AC that must be preserved by the NT to conclude noninferiority (Figure 10), using a previously described formula.[314] The %PE ranged from

0% (no different from placebo) to 100% (maximum effect of AC preserved). For absolute differences (e.g., percentages), the %PE was calculated as:

$$\%PE = (AC\ effect + noninferiority\ margin/AC\ effect)$$

For relative differences (e.g., hazard ratio), the %PE was calculated as:

$$\%PE = \log(AC\ effect \times noninferiority\ margin)/\log(AC\ effect)$$

The AC effect and noninferiority margin had to go in opposite directions. For example, a hypothetical 5% increased survival rate (AC vs placebo) and a -2.5% margin are compatible, just like a hypothetical hazard ratio of .8 (AC versus placebo) and margin of 1.2. If a %PE is less than 0% (i.e., negative percent), then the noninferiority margin is too wide and the NT is at risk of a “not inferior” conclusion, while actually being worse than placebo or another historical control.

Slightly modifying previous guidance,[313] we graded the quality of each noninferiority trial based on 4 criteria: (1) whether the margin was justified by previous data or clinical reasoning (yes vs no); (2) whether the selected margin could demonstrate that the NT preserves at least 50% of the AC effect (yes vs no/not capable of calculating %PE); (3) whether the type I error rate was consistent with the level of the confidence interval (yes vs no); and (4) how many analyses (e.g., intention to treat, per-protocol) were performed on the primary outcome (<2 or ≥2). Studies were graded as excellent (4/4 criteria), good (3/4), average (2/4), fair (1/4), or poor (0/4). The choice of 50% preserved AC effect was chosen since an excellent noninferiority trial achieves the goal of preserving a significant portion of the AC effect while also providing nonefficacy benefits.

The primary outcome of this investigation is the methodological quality of oncology noninferiority trials of antitumor or adjunct therapeutic interventions. For our quality assessment (using the 4 criteria), we report a sensitivity analysis excluding our %PE, owing to poor adherence. We calculated summary statistics using Google Sheets. No further statistical analyses were planned.

## *Results*

### *General Characteristics*

Our database search retrieved 337 articles, of which 110 were eventually included (Figure 11). These 110 articles were published most often in *Lancet Oncology* (n = 32), *Journal of Clinical Oncology* (n = 31), and *Annals of Oncology* (n = 17). The funding source was most often industry (n = 33), mixed (n = 19 with partial industry, n = 12 without partial industry), and public (n = 24). Protocols were available for 45/110 (40.9%) of noninferiority trials. Nonefficacy benefits of the NT were used as rationale in 88/110 (80.0%) trials, and a total of 103 nonefficacy benefits were cited. The most commonly cited NT nonefficacy benefits were fewer adverse events (n = 71), treatment convenience (n = 12), and cost (n = 10) (Table 9).

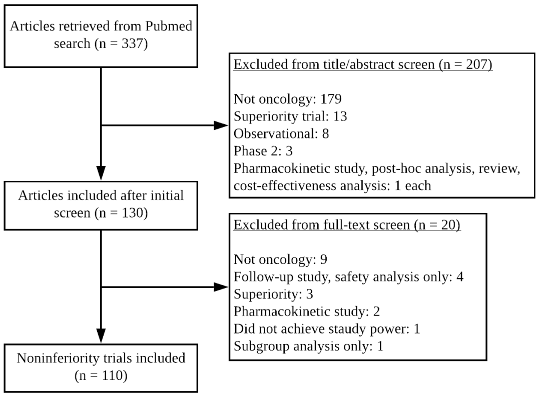
Articles retrieved from Pubmed search (n = 337)
Excluded from title/abstract screen (n = 207)
Not oncology: 179
Superiority trial: 13
Observational: 8
Phase 2: 3
Pharmacokinetic study, post-hoc analysis, review, cost-effectiveness analysis: 1 each
Articles included after initial screen (n = 130)
Excluded from full-text screen (n = 20)
Not oncology: 9
Follow-up study, safety analysis only: 4
Superiority: 3
Pharmacokinetic study: 2
Did not achieve staudy power: 1
Subgroup analysis only: 1
Noninferiority trials included (n = 110)

*Figure 11. Flow diagram of included studies.*

### *Trial Design*

In 18/110 (16.4%) noninferiority trials, the reported confidence interval and prespecified alpha value were not aligned. Authors most often used 1-sided alpha values (65/110, 59.1%), and the most common alpha value was .05 (34/110, 30.9%). Ten trials did not mention the chosen alpha level. No trials used a more favorable alpha level of greater than .05 (2-sided equivalent). Two-sided 95% confidence intervals were most often reported (78/110, 70.9%). Noninferiority trials most often prespecified 80% power (63/110, 57.3%). Primary endpoints were most often surrogate endpoints (e.g., PFS or response rate) (75/110, 68.2%) or OS (26/110, 23.6%). Hazard ratios were the most common outcome measure (73/110, 66.4%), followed by absolute differences (34/110, 30.9%).

| **Characteristic** | | | **No. (%)** |
|---|---|---|---|
| *Journal* | *Lancet: Oncology* | | 32 (29.1%) |
| | *Journal of Clinical Oncology* | | 31 (28.2%) |
| | *Annals of Oncology* | | 17 (15.5%) |
| | *The Lancet* | | 15 (13.6%) |
| | *New England Journal of Medicine* | | 12 (10.9%) |
| | *Jama Oncology; Cancer; Journal of National Cancer Institute* | | 1 (0.9%) each |
| *Funding* | Industry | | 33 (30.0%) |
| | Mixed | | 31 (28.2%) |
| | | Partial industry | 19 (17.3%) |
| | | No industry | 12 (10.9%) |
| | Public (e.g., government) | | 21.8%) |
| | Private (e.g., non-profit) | | 19 (17.3%) |
| | None | | 2 (1.8%) |
| | Not mentioned | | 1 (0.9%) |
| *Non-efficacy benefits** | Fewer adverse events | | 71 (68.9%) |
| | Convenience to patient (e.g., easier to administer) | | 12 (11.7%) |
| | Lower cost | | 10 (9.7%) |
| | Avoid future therapy (e.g., surgery) | | 3 (2.9%) |
| | Quality of life improvement* | | 3 (2.9%) |
| | Cosmetically better (e.g., for surgical procedures) | | 2 (1.9%) |
| | Optimize future therapy; Remove treatment delays | | 1 (1.0%) each |

*Table 9. Characteristics of included noninferiority trials (n = 110). * denominator of 103 for 103 total non-efficacy benefits cited; ** QoL coded for studies that mentioned QoL without specifics, e.g., no mention of QoL improvement by lowering adverse events*

Justification for the noninferiority margin was provided in 71.8% (79/110) of trials. Authors most often used previous data as justification for the chosen margin (n = 42), but only 40 trials cited a study containing such data. Despite 40 noninferiority trials citing a total of 73 potential studies as justification for the noninferiority margin, only 17 studies were eligible for calculation of %PE. Fifteen studies were included for calculation of %PE, with the remaining 2 eligible studies being passed over based on our decision algorithm (see Methods). Of the 15 noninferiority trials for which %PE could be calculated, 10 (66.7%) were designed to preserve greater than 50% AC effect, 4 (26.7%) were not designed to preserve 50% AC effect, and 1 (6.7%) could not calculated because the noninferiority hypothesis was to test a difference from zero using a special formula.[316] The median %PE was 56.8% (interquartile range 26.0%).

A total of 166 analyses were conducted for the noninferiority comparisons (e.g., 1 trial conducted intention-to-treat and per protocol). Half of the included trials (55/110, 50.0%) conducted only 1 analysis for the primary endpoint, with the most common analysis being intention to treat (48/55, 87.3%). Overall, the most common analysis was intention to treat (103/166, 62.0%), followed by per protocol (60/166, 36.1%). A total of 122 noninferiority comparisons (e.g., one NT versus AC) were made in the 110 included noninferiority trials. Authors most often concluded that the NT was not inferior to AC (70/110, 63.6%). The remaining noninferiority comparisons were either inferior (22/110, 20.0%), superior (10/110, 9.1%), or inconclusive (20/110, 18.2%) for the NT versus AC.

*Quality Judgment*

Seventy-seven (70.0%) of the 110 noninferiority trials included in the sample were scored as average (2/4 criteria; 51/110, 46.4%), fair (1/4 criteria; 22/110, 20.0%), or poor (0/4 criteria; 4/110, 3.6%) (Table 9). Only 5 (4.5%) noninferiority trials met all 4 quality criteria. Designing the noninferiority trial to preserve 50% of the AC effect was done the least often (10/110, 9.1%), with a failure to cite data to allow %PE calculations being the main driver of failing to meet this criterion. In a sensitivity analysis removing the %PE criterion, the number of excellent-quality noninferiority trials (all 3 remaining criteria) increased to 20 (20/110, 18.2%).

| **Criteria** | **Met [No. (%)]** | **Not met [No. (%)]** |
|---|---|---|
| *Margin to 50% preserved effect* | 10 (9.1%) | 100 (90.9%) |
| *>2 analyses (e.g., ITT and PP)* | 55 (50.0%) | 55 (50.0%) |
| *Matched alpha and confidence intervals* | 92 (83.6%) | 18 (16.4%) |
| *Use of clinical judgment or data to justify margin* | 74 (67.3%) | 36 (32.7%) |
| **Total of criteria met** | **No. (%)** | |
| *0/4* | 4 (3.6%) | |
| *1/4* | 22 (20.0%) | |
| *2/4* | 51 (46.4%) | |
| *3/4* | 28 (25.5%) | |
| *4/4* | 5 (4.6%) | |

*Table 10. Quality assessment of included noninferiority trials (n = 110).*

### *Discussion*

We found that oncology noninferiority trials are often of moderate to poor quality and often demonstrate key methodological shortcomings. These shortcomings include alpha values and confidence intervals that do not match, lack of citations for data that justify the chosen noninferiority margin, and prespecification of only 1 analysis (e.g., intention to treat only) for the primary endpoint. Altogether, these shortcomings are counterbalanced by the clearly delineated nonefficacy benefits expected from the NT and strong %PE in the 15 noninferiority trials in which calculating %PE was possible. However, the identified methodological shortcomings may lead to spurious conclusions of noninferiority that may be due to study design rather than the efficacy of the NT.

A previous study[304] showed that mismatched alpha values and confidence intervals (e.g., 2-sided 90% CI and .05 alpha) may result in spurious conclusions of noninferiority. In that study, recalculation using normal, more stringent confidence intervals (e.g., 95% instead of 90%) changed the conclusions of the noninferiority trial to be unfavorable to the NT. For example, if a 90% confidence interval is initially used and the authors conclude that the NT is not inferior to the AC by excluding the noninferiority margin, a 95% confidence interval may nullify this finding if the margin is included. Our reported rates of mismatched alpha values and confidence intervals may be cause for concern regarding the strength of oncology noninferiority conclusions—especially in the context of the clinical equipoise in sample size estimates and choice of noninferiority margin (i.e., possible sample size and margin as small as can be ethically justified) inherent to oncology noninferiority trials.[317] Moreover, the rate of noninferiority trials that only used intention-to-treat, which is known to bias a trial toward conclusions of noninferiority,[309,310,313] was concerning. Thus, for future oncology noninferiority trials, we recommend justifying an alpha value and matched confidence interval as well as both intention-to-treat and per protocol analyses in a publicly accessible study protocol.

Designing a noninferiority trial so that some of the AC effect versus placebo or another historical control is preserved is fundamentally important.[318] If the noninferiority margin is too large, a trial can conclude noninferiority while failing to preserve effectiveness over placebo or historical control (Figure 10). Thus, it is crucial that authors of noninferiority trials start with what effect the AC had over placebo or another control, and choose a clinically acceptable margin. We attempted to calculate the %PE by combining the noninferiority margin and AC effect against placebo or historical control into a formula previously described.[314] Unfortunately, only a small percentage of noninferiority trials cited studies that could be used for our calculation. We understand that it may be possible to infer expected AC efficacy based on trials of same-class drugs or of trials using the same AC for a different endpoint; however, we question the use of single-arm trials, observational studies, or studies in which no difference from placebo was found. Even using trials with different drugs, doses, or endpoints may introduce noise to the presumed AC effect that is used as the basis for the noninferiority margin. Encouragingly, however, where %PE could be calculated, the noninferiority trials were designed to preserve a median of 56.8% of the AC effect. In the future, we recommend that all noninferiority trials clearly delineate the justification for the noninferiority margin using data and citation where possible and with as much detail as possible if clinical judgment is all that is available.

This study has several key strengths and limitations. With regard to strengths, we used double data extraction to minimize bias in retrieved data. We also based our study on 3 previously published studies, but restricted our analysis to oncology noninferiority trials in a 6-year period. Thus, we believe our conclusions to be robust and relevant for the oncology community. With respect to limitations, we used very strict criteria for calculating %PE, which resulted in only a small subset of noninferiority trials being eligible for inclusion. It is likely that many noninferiority margins from our study not subjected to the %PE calculation were high quality, but we, and many readers, may be unable to confirm. Our inability to calculate %PE for all noninferiority trials also affected our quality assessment because 50%PE or more was one criterion. To remedy this limitation, we reported sensitivity analyses excluding this criterion.

In conclusion, we found that many oncology noninferiority trials clearly defined the expected nonefficacy benefits of an NT but exhibited some design shortcomings. We recommend addressing the following key methodological items in future noninferiority trials: (1) alpha values and confidence intervals that match, (2) prespecification of intention-to-treat and per protocol analyses for the primary endpoint, and (3) use of data, preferably with a citation, to justify a noninferiority margin that preserves a clinically meaningful effect of the NT compared to placebo or another historical control.

# CHAPTER X

# EVALUATION OF REPRODUCIBLE RESEARCH PRACTICES IN ONCOLOGY SYSTEMATIC REVIEWS WITH META-ANALYSES REFERENCED BY NATIONAL COMPREHENSIVE CANCER NETWORK GUIDELINES

***This work was previously published in* JAMA Oncology *with the following citation:***

Wayant C, Page MJ, Vassar M. Evaluation of Reproducible Research Practices in Oncology Systematic Reviews with Meta-analyses Referenced by National Comprehensive Cancer Network Guidelines. JAMA Oncol. Published online September 5, 2019. doi:10.1001/jamaoncol.2019.2564

## *Introduction*

Concerns are growing about the reproducibility of biomedical research.[319,320] Many of these concerns stem from research practices that lack transparency, including poor reporting of study methodology[321] and failing to make study data publicly available.[143] As a result, efforts to reproduce biomedical research findings have been thwarted.[322,323] The vast majority of efforts to reproduce research findings have been dedicated to primary studies, such as clinical trials, and little effort has been dedicated to reproduce higher levels of evidence, such as SRs. The first studies to holistically evaluate the reproducibility of SRs and meta-analyses in the biomedical literature found that authors frequently fail to employ reproducible research practices.[143,324] However, only a small proportion of the SRs evaluated in previous investigations were for oncology interventions, leaving unanswered questions for researchers in this field, oncologists, and policy makers.

For this investigation of the reproducibility of oncology SRs, we identified SRs cited in NCCN CPGs. The NCCN set of guidelines are one of many available to oncologists; however, a survey of oncologists showed that NCCN guidelines were more likely to influence clinical practice than other popular oncology guidelines[325]. Further, NCCN guidelines cover all blood and solid cancers, thus making them ideal for a broad investigation such as this. The primary objective of this investigation is to evaluate the reproducibility of meta-analyses in oncology SRs cited by the 49 NCCN guidelines for the treatment of cancer by site. The secondary objective is to evaluate whether Cochrane reviews or SRs that report adherence to PRISMA employ more reproducible research practices.

### *Methods*

The protocol for this investigation is publicly available via the Open Science Framework[326]. We defined an SR according to the PRISMA for protocols definition: articles that explicitly stated methods to identify studies (i.e., a search strategy), explicitly stated methods of study selection (e.g., eligibility criteria and selection process), and explicitly described methods of synthesis (or other type of summary).[327] Since NCCN guidelines update regularly throughout each year, all guidelines were manually downloaded as PDFs on May 6, 2018 to avoid citations being added to the guideline during the course of our investigation.[328] To identify SRs we manually screened the reference lists and Discussion narratives of all NCCN CPGs for the treatment of cancer. We extracted all references with “systematic review”, “meta-analysis”, “metaanalysis”, and any references without the keywords in the title that are discussed as an SR or meta-analysis by guideline authors. We also extracted any cited references that were published in the Cochrane Database for Systematic Reviews. All extracted references were added to a PubMed collection and exported to Rayyan[256] for title and abstract screening.

We screened articles using the liberal acceleration method whereby one author (CW) was required to mark a record for inclusion and two authors (CW, MP) were required to mark a record for exclusion. Next, two authors (CW, MP) screened the full-text of potentially relevant articles for inclusion. Key inclusion criteria were SRs published in 2011 or later with at least one meta-analysis that included at least one randomized-controlled trial. We chose to include only SRs published after 2011 to allow time for uptake of the 2009 PRISMA Statement. Thus, all included SRs are accountable to currently accepted reporting quality standards. SRs of individual patient data or of primary studies other than clinical trials, network meta-analyses, and pooled analyses of clinical trials were excluded.

To extract data for this study we developed a pilot-tested Google Form based on the extraction form used in a similar, previous study[143]. Extracted data items were related to the number of meta-analyses reported, reporting of summary statistics for each individual study, use of fixed-effect versus random-effect models, interpretation of tests of heterogeneity and small-study effects, and types of subgroup and sensitivity analyses performed. We extracted data for all meta-analyses, but certain items were dedicated to the index meta-analysis, which we defined as the primary meta-analysis for the primary endpoint. If there was no primary endpoint mentioned, we used the first reported meta-analysis as the index meta-analysis and inferred the primary endpoint from there. We counted meta-analyses by summing the number of summary effects in forest plots, written narrative, and supplemental appendices. Duplicate meta-analytic effects were only counted once. We counted subgroup effects that were derived from an analysis of at least two studies, as well as the overall summary effect that synthesized all subgroup effects. We only counted sensitivity analyses that were expressly described with a summary effect in the paper or the supplemental material.

To be considered reproducible in theory an analysis must have three elements: 1) effect estimate and measure of precision (e.g., hazard ratio with 95% confidence interval); 2) clear list of studies included for each analysis; 3) for subgroup and sensitivity analyses, it was clear which studies were included in each group or level.

Data from all SRs were extracted by CW. A random sample of 15% of the included SRs was extracted in duplicate by MP. MV adjudicated discrepancies in the double-extracted 15% sample. Any item that had at least one discrepancy was reviewed a second time in the 85% of other studies by CW. A complete list of items with a discrepancy are available, along with our protocol and data, via the Open Science Framework[326].

Summary statistics and measures of central tendency (e.g., median with interquartile range (IQR)) were calculated using Microsoft Excel. We planned to use STATA 15.1 to calculate risk ratios and 95% confidence intervals for the comparisons between Cochrane and non-Cochrane SRs, and between SRs self-reporting use of PRISMA versus not, but owing to disparate numbers of Cochrane and non-Cochrane SRs, we only report the comparisons of SRs stratified by PRISMA adherence and year of publication. We conducted sensitivity analyses for meta-analyses presented in figures and for those published as supplementary material to investigate potential factors contributing to reproducibility.

### *Results*

*Characteristics*

We identified 1,124 potential SRs from our survey of the 49 NCCN guidelines for the treatment of cancer by site. Five NCCN guidelines did not cite any SRs. An additional 19 CPGs did not have any SRs that met the inclusion criteria. After removing duplicates and screening all articles, 154 SRs were included from 25 guidelines (Figure 12)[326]. There was high agreement between reviewers (94.0%) for studies extracted in duplicate.

Half of the included SRs were either a Cochrane review or mentioned adherence to PRISMA (77/154, 50.0%). Eighteen (11.7%) SRs were Cochrane SRs, and 60 (39.0%) adhered to PRISMA. Of the SRs that received funding, public sources (e.g., government) were most common (36/78, 46.2%). The SRs included a median of 14 (IQR 7.25 - 29.75) meta-analytic effect estimates, including those from subgroup and sensitivity analysis. Additional characteristics of our sample are reported in Table 17.

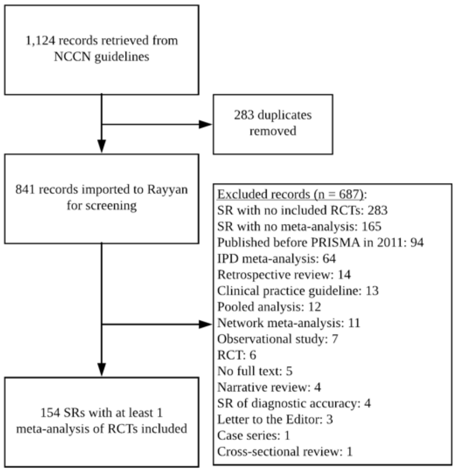
1,124 records retrieved from NCCN guidelines
283 duplicates removed
841 records imported to Rayyan for screening
Excluded records (n = 687):
SR with no included RCTs: 283
SR with no meta-analysis: 165
Published before PRISMA in 2011: 94
IPD meta-analysis: 64
Retrospective review: 14
Clinical practice guideline: 13
Pooled analysis: 12
Network meta-analysis: 11
Observational study: 7
RCT: 6
No full text: 5
Narrative review: 4
SR of diagnostic accuracy: 4
Letter to the Editor: 3
Case series: 1
Cross-sectional review: 1
154 SRs with at least 1 meta-analysis of RCTs included

*Figure 12. Flow diagram of included studies.*

Only 88 (57.1%) SRs labelled their primary endpoint (Table 18). Thus, we inferred the primary endpoint in the remaining 66 SRs from the index (first reported) meta-analysis. Seventy-three (47.4%) primary endpoints were all-cause mortality. A median of 8 (IQR 5 - 12) primary studies with a median 1,914 (IQR 917 - 3,941) patients was included in each index meta-analysis. Seventy-nine (51.2%) index meta-analyses included a subgroup analysis and 54 (35.1%) included a sensitivity analysis.

*Reproducible research practices: Overall*

There was a total of 3,696 meta-analytic effect estimates, including subgroup and sensitivity analyses in the 154 SRs, but only 2,375 (64.3%) were reproducible in theory. All meta-analyses were reproducible in theory in 100 (64.9%) SRs, and in 139 (90.3%) SRs there was at least one meta-analysis that could potentially be reproduced. Summary statistics (e.g., event rates) for studies included in the index meta-analysis were reported in 107 (69.5%) SRs, but only 39 (25.3%) mentioned whether or not missing data was imputed and included in the index meta-analysis. Missing data was reported to have been imputed in 29 of these 39 SRs, but it was not clear which exact data points were imputed in all 29. Similarly, only 29 SRs mentioned whether unpublished data were retrieved from primary study authors, with 17 affirming that authors were contacted. However, only 3/17 (17.6%) were clear about which data were retrieved.

Eighty-seven (56.5%) SRs generated funnel plots to assess for publication bias, but only 49/87 (56.3%) presented the funnel plot in the SR or supplemental appendix.  In many SRs the number of studies included in the funnel plot was unclear (28/87, 32.2%). Only 62 SRs cited the guide they used to interpret their $I^2$ statistic, with the most common guide being by Higgins, et al[329]. Sixty-one (39.6%) authors decided between a random- or fixed-effects model based on the statistical heterogeneity of the included studies, but 31/61 (50.8%) did not report the amount of heterogeneity necessary to use a random-effects model.

Random-effects models were used for 91/154 (59.1%) index meta-analyses, but specific information about the between-trial variance estimator (e.g., Dersimonian and Laird[330]) were not reported in 45/91 (49.5%). Subgroup analyses were included in 79/154 (51.3%) index meta-analyses, but only 51/79 (64.6%) were fully reproducible in theory. Of the 54 sensitivity analyses that accompanied index meta-analyses, only 34 (63.0%) were fully reproducible in theory. Only 1 SR — a Cochrane SR — included a link to an online dataset.

When considering only the 2,341 of 3,696 meta-analyses that were presented on forest plots, we determined that 2,195/2,341 (93.7%) were reproducible in theory, since they included numerical point estimates (or event rates conducive to calculating point estimates) and a list of included studies. Compared to meta-analyses not published in figures (180/1,355), forest-plot-based meta-analyses were more often reproducible in theory (uRR 8.4; 95% CI, 7.2 - 9.7). When considering only meta-analyses published as supplemental material, we determined that 368/642 (57.3%) were reproducible in theory. Compared to main-text meta-analyses (2,007/3,054), supplemental meta-analyses were less often reproducible in theory (uRR .74; 95% CI .65 - .86). Both sensitivity analyses are unadjusted and should be interpreted with caution, especially the supplemental versus main-text analysis, which is likely confounded by forest-plot-based meta-analyses.

| **Reproducible research practice** | | **All (n = 154)** | **PRISMA (n = 60)** | **non-PRISMA (n = 94)** |
|---|---|---|---|---|
| *Reported the data needed to recreate all meta-analytic effect estimates in the SR* | | 100 (64.9%) | 36 (60.0%) | 64 (68.1%) |
| *Reported the data needed to recreate the index meta-analytic effect estimate* | | 140 (90.9%) | 58 (96.7%) | 82 (87.2%) |
| *Reported summary statistics for each individual study in the index meta-analysis* | | 107 (69.5%) | 42 (70.0%) | 65 (69.1%) |
| *Reported effect estimates and measures of precision for each individual study in the index meta-analysis* | | 140 (90.9%) | 58 (96.7%) | 82 (87.2%) |
| *Reported that some data in the index meta-analysis had been imputed* | | 39 (25.3%) | 14 (23.2%) | 25 (26.6%) |
| | Clear which data were imputed and how | 0 (0.0%) | 0 (0.0%) | 0 (0.0%) |
| *Reported that some data in the index meta-analysis had been obtained from the study author/sponsor* | | 17 (11.0%) | 7 (11.7%) | 10 (10.6%) |
| | Clear which data were obtained | 3 (1.9%) | 0 (0.0%) | 3 (3.2%) |
| *Reported (or inferred) the type of random-effects method used for the index meta-analysis* | | 78/91 (85.7%) | 22/43 (51.2%) | 24/48 (50.0%) |
| *Reported the data needed to recreate each subgroup analysis for the index meta-analysis* | | | | |
| | For all subgroup analyses | 51/79 (64.6%) | 26/40 (65.0%) | 25/39 (64.1%) |
| | For some subgroup analyses | 1/79 (1.3%) | 0/40 (0.0%) | 1/39 (2.6%) |
| | Not for any subgroup analysis | 27/79 (34.2%) | 14/40 (35.0%) | 13/39 (33.3%) |
| *Reported the data needed to recreate each sensitivity analysis for the index meta-analysis* | | | | |
| | For all sensitivity analyses | 34/54 (63.0%) | 12/23 (52.2%) | 22/31 (71.0%) |
| | For some sensitivity analyses | 2/54 (3.7%) | 2/23 (8.7%) | 0/31 (0.0%) |
| | Not for any sensitivity analysis | 18/54 (33.3%) | 9/23 (39.1%) | 9/31 (29.0%) |
| *Mention of access to data sets and statistical analysis code used to perform analyses* | | 1 (0.6%) | 0 (0.0%) | 1 (1.1%) |

*Table 11. Reproducible research practices of systematic reviews that underpin the National Comprehensive Cancer Network clinical practice guidelines for the treatment of cancer by site.*

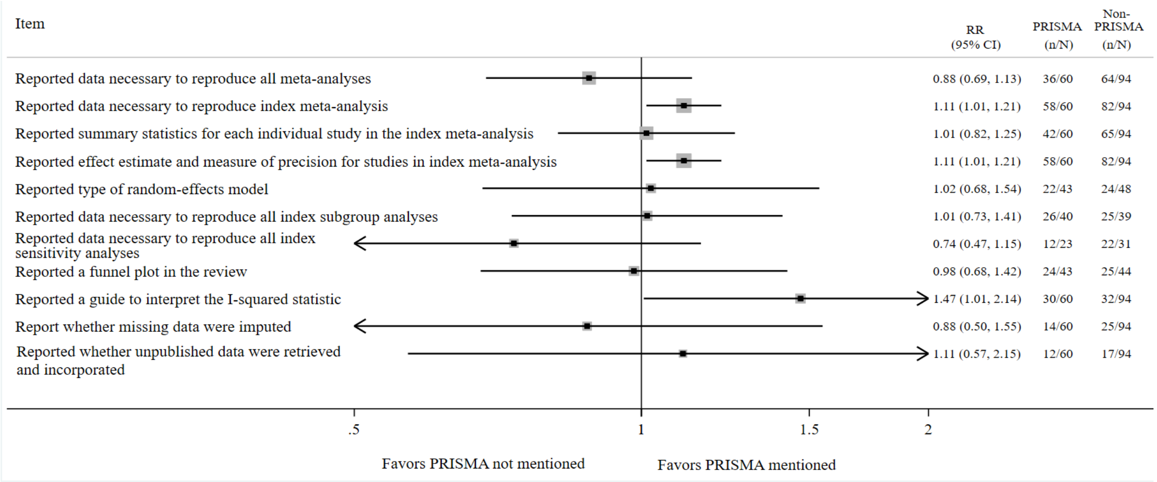

*Figure 13. Comparison of reproducible research practices in systematic reviews that did and did not report adherence to PRISMA guidelines.*

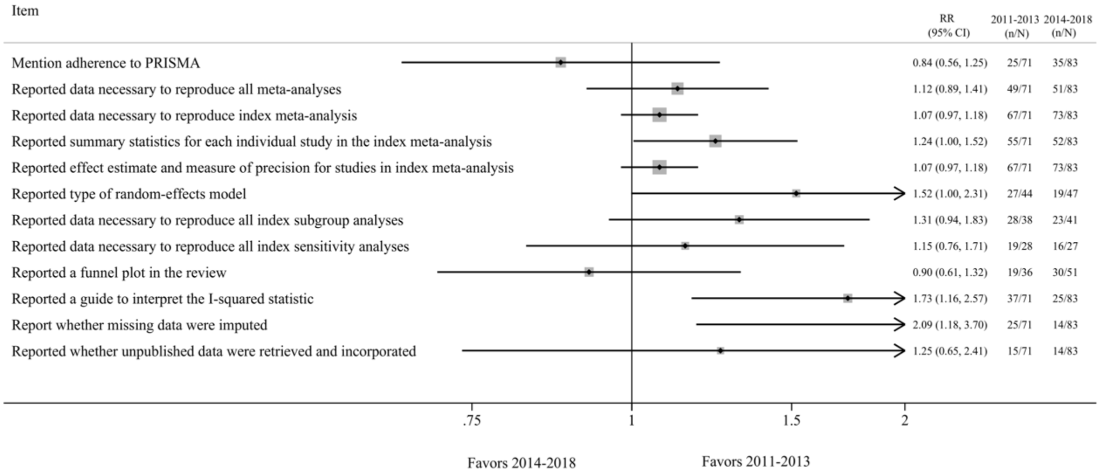

*Figure 14. Comparison of reproducible research practices in systematic reviews before and after uptake of PRISMA guidelines.*

### *Stratified analyses of reproducible research practices*

We limit our analysis of Cochrane and non-Cochrane reviews to summary statistics owing to large differences in group sample sizes. One of 18 (5.6%) Cochrane SRs and 59 of 136 (43.4%) non-Cochrane SRs stated that they adhered to PRISMA guidelines. In 16/18 (88.9%) Cochrane SRs all included meta-analyses were reproducible in theory compared to 85/136 (62.5%) non-Cochrane SRs. Regarding sensitivity and subgroup analyses, all were reproducible in theory in Cochrane SRs. In non-Cochrane SRs only 29/48 (60.4%) with sensitivity analyses and 49/77 (63.6%) with subgroup analyses provided enough information to make these analyses reproducible. All data for comparisons between SRs that did and did not mention PRISMA are in Table 11 and Figure 13. Data for our analysis by year of publication is shown in Figure 14.

### *Discussion*

The results of our investigation demonstrate that reproducible research practices are commonly implemented for primary analyses, but far less so for secondary, subgroup, and sensitivity analyses. Moreover, figure-based (e.g., forest plot) meta-analyses were far more reproducible than other meta-analyses, and our sensitivity analysis shows that the main driver of whether a meta-analysis was reproducible or not was based on it being published in a forest-plot or not. SRs cited by oncology practice guidelines may represent the most important cohort of oncology SRs, since these SRs inform guideline recommendations, in some cases. Yet, despite recent improvements in the quality of SRs after the publication of the PRISMA statement[331], we found that key items were missing from oncology meta-analyses, which may hinder their reproducibility. The ability to reproduce all meta-analytic effects — even for secondary endpoints, since SRs are not powered for one endpoint like clinical trials — is fundamentally important, since scientific progress requires trustworthy results. And while the inability to reproduce study findings does not mean the study findings are false, it may affect the interpretation of results, especially since our study defined "reproducibility" for main effects as the reporting of a summary effect, measure of precision, and list of included studies.

Our findings are comparable to those from a previous, similar study that examined the reproducible research practices of a cross-section of SRs and meta-analyses that were published in February of 2014[143]. That study found that 73% of meta-analytic effects were reproducible in theory, compared to the 64.3% found in our study. For articles in this study, adhering to PRISMA and citing a guide to interpret statistical heterogeneity both seemed to improve the reporting of effect estimates and measures of precision for the index meta-analysis. These effects are either small or imprecise and should be interpreted accordingly.

This study has several key strengths and limitations. Our sample of 154 is 40% larger than the previous study of reproducible research practices and is focused on only one area of medicine. Unlike previous investigations of data reporting in SRs[43,332–336], we extracted whether data necessary to reproduce meta-analyses (e.g., summary statistics or effect estimates) were available from published reports, and whether subgroups or sensitivity analyses differed from the index meta-analyses in this regard. Concerning limitations, our sample of SRs may not be generalizable to all SRs of oncology interventions, because we relied on the citations in NCCN guidelines. It is possible that other specialized organizations (e.g., American Society of Hematology for blood cancers) cite different SRs. Further, it may be that other SRs of oncology interventions are more or less reproducible in theory than those in this study. We used double data extraction for only 15% of the included studies, which may increase the chance of data extraction errors. Despite high percent agreement between authors, in order to mitigate the possibility of these errors, we extracted data a second time for all items with a discrepancy and used a third-party adjudicator. These quality checks are consistent with previous studies[143,337]. Further, the absence of data to reproduce a meta-analysis effect does not necessarily imply it was incorrectly estimated, only that the availability of the data to reproduce may improve confidence for some readers in its accuracy.

In conclusion, we recommend that SR authors incorporate more reproducible research practices and expect guideline authors to evaluate whether existing SRs are reproducible. We further recommend journals encourage authors to present all meta-analyses in figures, since standard graphical output for meta-analyses in most statistical packages includes a list of included studies and numerical point estimates. In this study, these two items alone were necessary to reproduce a summary effect, in theory. A guideline development group may downgrade the quality of SR data if they feel that the findings are not trustworthy. We further recommend earnest adherence to PRISMA, since many reproducible research practices we investigated are addressed therein, indicating that authors may incompletely adhere to PRISMA recommendations. Authors should make use of data repositories, such as the Open Science Framework, to store data, supplemental material, or other necessary items that ensures the reproducibility of findings.

# CHAPTER XI

# RISK OF BIAS AND QUALITY OF REPORTING IN COLON AND RECTAL CANCER SYSTEMATIC REVIEWS CITED BY NATIONAL COMPREHENSIVE CANCER NETWORK GUIDELINES

***This work was previously published in the* Journal of General Internal Medicine *with the following citation:***

Wayant C, Puljak L, Bibens M, Vassar M. Risk of Bias and Quality of Reporting in Colon and Rectal Cancer Systematic Reviews Cited by National Comprehensive Cancer Network Guidelines. J Gen Intern Med. Published online January 16, 2020. doi:10.1007/s11606-020-05639-y

---

## *Introduction*

SRs combine results from similar, individual studies in an attempt to provide a reliable answer to a healthcare question[338]. Previous SRs have demonstrated benefit to both physicians and patients. An iconic example of how SRs have influenced clinical practice concerns antenatal corticosteroid use in women at risk for preterm birth[71]. This SR demonstrated a survival benefit for preterm infants and resolved unanswered clinical questions, such as the long-term effects of corticosteroids on surviving infants. The authors of this SR reported their methodology and conducted their study in a manner that promotes reproducibility and trustworthiness. Examples of such practices include publishing the search strategy used to identify included studies, assessing the included studies for risk of bias, and using robust statistical methods to combine these studies for determining the pooled treatment effect. These practices, however, are not common as previous studies suggest that SRs often fail to report detailed search strategies or evaluate for risk of bias using valid tools[337,339]. Such incomplete SR methodology may lead to biased results, the consequences of which are far-reaching, including spurious alterations to clinical practice and future research questions. These consequences are especially harmful if the SR is cited to support CPG)recommendations.

CPGs are consensus documents developed by a group of experts that are designed to guide patient care[88]. SRs are often used, alongside robust clinical trials, to assign level 1 evidence to CPG recommendations[340]. However, for an SR to be a trustworthy and accurate source of clinical information, its methods and reporting must first be robust. Previous investigations of SRs underpinning CPG recommendations have identified suboptimal methodology and reporting[43,333,334]. Such SRs may be irreproducible, and the critical appraisal of their summary effects by CPG development groups may be compromised.

SRs are also critically important to the prevention, diagnosis, and treatment of colon and rectal cancer. Currently, colorectal cancer is being diagnosed in patients under age 50 at an increasing rate[341]. Even worse, there is evidence that colon cancer in younger patients may differ from colon cancer in older adults with respect to clinical presentation, pathologic findings, and tumor biology[342]. Therefore, there is a fundamental need for robust research based on rigorous methodology to continue the advancements in understanding preventing, diagnosing, and treating colorectal cancer. SRs are likely to play a key role in these advancements.

Therefore, the primary objective of this study is to assess the risk of bias and reporting quality in SRs cited in the NCCN guidelines for the treatment of colon (Version 2.2018)[343] and rectal (Version 1.2018)[344] cancer, since NCCN guidelines are heavily used by physicians to guide patient care[325] and SRs are the highest level of medical evidence. To do so, we applied the novel Risk of Bias In Systematic Reviews (ROBIS) tool[100] and Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) checklist[345].

### *Methods*

In this review, we adhered to PRISMA guidance where possible and applicable[327], despite this study not being an SR. Our choice to adhere to PRISMA was made because no validated reporting checklist for cross-sectional or meta-epidemiological studies exists. We defined an SR according to the PRISMA-P (PRISMA for protocols) definition: articles that explicitly stated methods to identify studies (i.e., a search strategy), explicitly stated methods of study selection (e.g., eligibility criteria and selection process), and explicitly described methods of synthesis (or other type of summary)[327]. SRs were gathered from a previous study whose protocol is available via the Open Science Framework[326]. To identify SRs in the previous study, one author (CW) manually screened the reference list of and Discussion (main narrative) section of the NCCN colon and rectal cancer guidelines. Keywords searches were conducted for studies referenced as a “systematic review”, “meta-analysis”, “review”, or “metaanalysis”. Any referenced papers published in the Cochrane Database for Systematic Reviews were also extracted. All extracted references were added to a PubMed collection and exported to Rayyan[256] for title and abstract screening. In accordance with the previous study from which our sample is derived, an SR was included if it had at least one meta-analysis that included at least one randomized-controlled trial. Further, each SR that met these criteria must have been published after 2011 to allow uptake of the PRISMA statement (published in 2009). To extract data for this study we developed Google Forms based on the ROBIS and PRISMA statements. Two authors extracted data in duplicate with masking for ROBIS (CW, LP) and for PRISMA (CW, MB). All discrepancies were resolved between the authors, with the availability of a third-party adjudicator (MV).

The ROBIS statement assesses whether an SR is at risk of bias based on its methods and conduct. ROBIS includes 3 phases: 1) assess relevance (optional), 2) identify concerns with the SR process, and 3) judge risk of bias of the SR. We opted to exclude the first phase of assessing relevance, since all SRs from the NCCN colorectal guidelines are relevant to our research question. In the second phase, signaling questions are asked to guide an investigator through 4 bias domains: 1) study eligibility criteria, 2) identification and selection of studies, 3) data collection and study appraisal, and 4) synthesis and findings. Signaling questions are answered as “yes”, “probably yes”, “no”, “probably no”, and “no information”. We followed the guidance of the ROBIS statement manual when answering signaling questions. Based on the answers to the signaling questions in each domain, each domain is assigned a risk of bias grade. Potential grades include “high”, “low”, or "unclear". In the third phase, signaling questions are again asked, except these questions relate to the overall reliability of the SR findings. If limitations are identified in phase 2, SR authors will be required to address these limitations and interpret the findings accordingly. Further, SR authors will be required to not emphasize findings based on statistical significance alone. After completing phase 3, a summary judgement (e.g., high, low, or unclear) regarding the risk of bias for the SR will be rendered. For this study, we distinguished between high and unclear risk of bias based on the completeness of SR reporting. For example, to be judged high risk of bias, the SR would have to report the use of flawed methods, such as a flawed risk of bias scale, use of only one database to gather studies, or single-author data extraction. If an SR did not report enough information for us to determine whether the methods were at high or low quality, we judged that SR with unclear risk of bias.

Contrary to ROBIS, PRISMA assesses reporting quality, so rather than asking if an item was conducted adequately, PRISMA asks whether an item was reported. For example, whereas ROBIS may ask about the adequacy and comprehensiveness of the search strategy, PRISMA asks if a search strategy was reported. This distinction is important for complementing the assessment of risk of bias in SR methodology measured using ROBIS. The PRISMA checklist contains 27 items divided into 7 domains: Title, Abstract, Introduction, Methods, Results, Discussion, and Funding. For each item, we judged whether an SR fully reported what was required by PRISMA and scored each item with a 1 (Yes – fully reported) or 0 (Not reported or Partially Reported). Our rationale for partial reported being grouped with "not reported" is that PRISMA does only asks whether an item is mentioned, not that it was methodologically robust. So, failure to completely report an item indicates that a key piece of that item is not available to readers. For example, Item 5 requires SR authors to indicate if a protocol exists and direct readers to it with a citation or registration number. Failure to direct readers means readers are unable to access the protocol, just as if the SR authors did not mention a protocol at all. After scoring each PRISMA item, we summed the adherence across each article and each item. It should be noted that PRISMA is not a measurement tool, but a reporting checklist. Despite that fact, PRISMA has been used in numerous previous studies as a measurement tool, since no other validated option to assess reporting quality exists.

We used Google Sheets for all summary statistics and measures of central tendency (medians and interquartile ranges (IQR)).

### *Results*

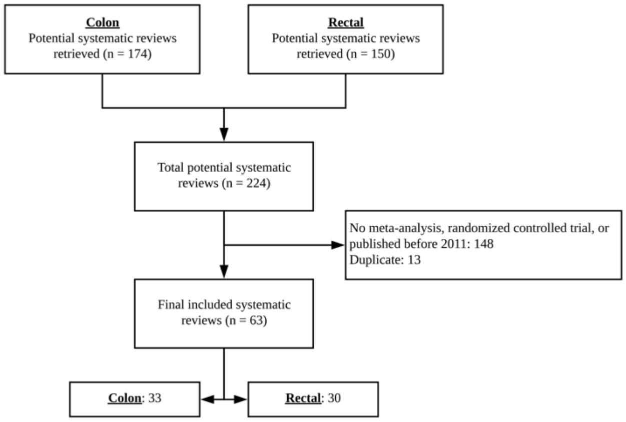
Colon
Potential systematic reviews retrieved (n = 174)
Rectal
Potential systematic reviews retrieved (n = 150)
Total potential systematic reviews (n = 224)
No meta-analysis, randomized controlled trial, or published before 2011: 148
Duplicate: 13
Final included systematic reviews (n = 63)
Colon: 33
Rectal: 30

*Figure 15. Flow diagram of included and excluded studies.*

Sixty-three SRs (33 Colon, 30 Rectal) were included in this study (Figure 15). The 63 SRs included a median of 10 (IQR 7-16) studies and a median of 3,160 (IQR 1,270-5,825) patients. Twenty-four (38.1%) SRs stated that they adhered to PRISMA guidelines. The included SRs were cited a total of 76 times, overwhelmingly for support of NCCN committee recommendations (56/76, 73.7%). SRs were also cited as evidence of harm for available therapies (10/76, 13.2%), as evidence that contradicts the committee recommendations (5/76, 6.6%), and as background evidence where no recommendation was given (5/76, 6.6%). All primary data and the protocol from this investigation are available via the Open Science Framework[346].

Using ROBIS, only 3 (4.8%) SRs were judged with low risk of bias, 35 (55.6%) SRs were judged with unclear risk of bias, and 25 (39.7%) SRs were judged with high risk of bias (Table 12). Across all SRs, the individual bias domains at the highest risk of bias, were domains 1 (protocol and eligibility criteria) and 2 (methods to identify and select studies). Twenty-eight (44.4%) SRs were at high risk of bias for domain 1 and 26 (41.3%) were at high risk of bias for domain 2. Specific areas of concern in these two domains were the lack of information about publication of an SR protocol, language restrictions, choice of bibliographic databases, and searches for grey literature. Domains 3 (data collection and appraisal) and 4 (synthesis of findings) were predominantly judged as unclear risk of bias, corresponding to a frequent lack of critical information that would have aided our assessments. Individual study scores are shown in Table 19.

| | **DOMAIN 1**<br>Protocol &<br>Eligibility Criteria | **DOMAIN 2**<br>Methods to identify<br>and/or select studies | **DOMAIN 3**<br>Data collection &<br>Appraisal | **DOMAIN 4**<br>Synthesis of<br>findings | **Consensus** |
|---|---|---|---|---|---|
| *Low* | 6 | 5 | 5 | 4 | 3 |
| *Unclear* | 29 | 32 | 43 | 46 | 35 |
| *High* | 28 | 26 | 15 | 13 | 25 |

*Table 12. Summary of risk of bias judgments for included systematic reviews (n = 63)*

Across all studies, the median adherence to PRISMA was 74.1% (IQR 69.2%-80.0%), corresponding to approximately 20 of 27 items (Table 20). Two items had 100% adherence: Item 3 (rationale for SR) and Item 21 (presentation of results with measures of precision). Thirteen additional items had adherence greater than 75%, with 7 items maintaining adherence greater than 90%. Only 3 items had adherence lower than 25%: Item 8 (search strategy), Item 5 (protocol and registration), and Item 4 (provision of PICO-format research question). There was no difference between SRs that adhered to PRISMA (n = 24) and did not adhere to PRISMA (n = 39) in terms of number of fully reported items (20 PRISMA vs. 20.5 no PRISMA).

### *Discussion*

Our investigation found that SRs cited in colorectal guidelines are frequently at unclear or high risk of bias and do not report key SR items that are important for the critical appraisal of results. Specifically, that our predominant risk of bias judgement was unclear, signals that much of the critical SR methodological items were missing or poorly described. Our finding — that SRs adhered to a median of 20/27 PRISMA items — may appear at odds with our risk of bias findings. However, the difference in these two findings highlights our key takeaway: an SR item may be reported but still represent a flawed method, thus placing the SR at risk of bias. Thus, our findings identify two key action items for future and ongoing SRs in colorectal cancer: ensure SRs report all items from PRISMA and ensure SRs describe methods in enough detail to facilitate critical appraisal of results.

Two key examples of how missing or poorly described information may affect the critical appraisal of an SR relate to study protocols and risk of bias evaluations. In our sample, SRs rarely directed the reader to a publicly available, *a priori* protocol (2/63, 3.2%). It has been shown that SRs, like randomized-controlled trials[36,136], exhibit significant rates of selective outcome reporting — defined as the selective inclusion, omission, or alteration of study outcomes, often due to statistical significance[291]. Thus, the lack of a publicly available protocol leaves the possibility that SR results are published at the author's discretion, rather than at the behest of a prespecified protocol. Similarly, a lack of detail regarding risk of bias evaluations may compromise the validity of meta-analytic effects in an SR. In our study, authors often reported that a risk of bias evaluation was conducted (46/63, 73.0%), but further inspection of the risk of bias methods showed that many authors used outdated, flawed tools. For example, authors frequently used the Jadad scale for assessing risk of bias of included clinical trials. The Jadad scale is notorious for its omission of allocation concealment as a bias domain, and according to the Cochrane Handbook, use of the Jadad scale is "explicitly discouraged"[347]. Thus, the use of the Jadad scale leaves the possibility that interventional effects shown in the included colorectal SRs are confounded by bias that is undetected by SR authors. Furthermore, even if authors used Cochrane risk of bias tool, they often reported only judgment for individual risk of bias domains, without an accompanying comment that explained the judgment. It has been shown previously that authors frequently make erroneous judgments (i.e. judgments that were not in line with the accompanying comment) and thus, not in line with recommendations available in the Cochrane Handbook[348–350]. Therefore, inadequate reporting of Cochrane risk of bias tool prevents readers to verify accuracy of authors' judgments.

The cohort of SRs we analyzed are unique since these SRs informed the evidence base of NCCN colorectal guidelines. However, this sample of SRs is likely not the only, or even the primary, source of evidence for most NCCN recommendations, since the field of oncology relies heavily on randomized-controlled trial data. Indeed, the NCCN categories of recommendations simply state that "high-level evidence" and "uniform NCCN consensus" is necessary to achieve Level 1 evidence status[351]. Nonetheless, the findings from our study warrant concern due to the predominance of unclear or high risk of bias judgements and variability in reporting quality. For example, in the NCCN rectal cancer guidelines, seven SRs were cited in the discussion of laparoscopy vs. open resection[352–358]. Five of these SRs were at high or unclear risk of bias, while 2 were at low risk, including the only Cochrane review. There was no discussion of the risk of bias for any of these SRs. This oversight may be reasonable in this case because of the dearth of other data available and cited for laparoscopy, all pointing to a fairly certain conclusion of its risks and benefits. Moreover, in this case the low risk of bias SRs had similar findings as the high and unclear risk of bias SRs. However, even this scenario highlights an important point — risk of bias assessments are crucial to reasoned discussions and serve to augment the ongoing, skillful clinical appraisal inherent to CPG panel discussions. In this case, where the benefits and risks of laparoscopy are fairly well-established, the harm of omitting risk of bias from a CPG discussion may be benign, but for emerging therapies with less certain benefit, risk of bias evaluations are necessary because the risk of false positive or negative results may have a broad impact of CPG recommendations and clinical practice.

This study has several key limitations. First, our findings may not be generalizable to all colorectal SRs, since we only evaluated SRs cited by the NCCN rather than all colorectal SRs available. Next, we discourage the interpretation of our findings to mean that NCCN recommendations are at risk of bias, since the NCCN recommendations rely on other robust research, such as clinical trials, that we did not include in our investigation. Any judgments about the quality of NCCN recommendations would need to be supported with thorough assessment of all evidence included and validated tools for assessment of clinical guidelines. Moreover, the included NCCN guidelines included 1,698 total references, so our 63 included SRs represents only a small fraction of the cited evidence. Finally, this study is limited by investigating only guidelines written for healthcare professionals, rather than NCCN guidelines for patients. In conclusion, our investigation of the risk of bias and quality of reporting of SRs referenced by the NCCN guidelines for colon and rectal cancer found that SRs are commonly at high risk of bias and do not fully report key items. Specifically, we found that an SR item may be mentioned, but may report a flawed method or incompletely report all aspects of the item. The implication for the treatment and management of colon and rectal cancer, which relies on high-quality evidence for demographically diverse patients, is that summary effects may not exemplify the trust normally imputed on SRs and meta-analyses. Further, even though the objective of our investigation is not to question the strength of NCCN guideline recommendations, our findings may be of concern to oncologists who heavily rely on NCCN recommendations. The NCCN developers use what literature is available to formulate recommendations, and thus, we recommend more stringent SR methodology and reporting be enforced in journal publications. When readers or guideline developers encounter a biased SR, we recommend careful critical appraisal of the results and conclusions, since bias may result in false positive or false negative results.

# CHAPTER XII

## EVALUATION OF THE NCCN GUIDELINES USING THE RIGHT STATEMENT AND AGREE-II INSTRUMENT: A CROSS SECTIONAL REVIEW

***This work was previously published in the* British Medical Journal: Evidence Based Medicine *with the following citation:***

Wayant C, Cooper C, Turner D'arcy, Vassar M. Evaluation of the NCCN guidelines using the RIGHT Statement and AGREE-II instrument: a cross-sectional review. BMJ Evid Based Med. 2019;24(6):219-226.

### *Introduction*

Robust, clearly reported CPGs are essential for evidence-based clinical practice. The Institute of Medicine recognizes CPGs as necessary reference material for physicians seeking to optimize patient care.[88] CPGs are capable of increasing the quality of patient care and improving patient outcomes[359], but the adoption of low-quality guidelines may result in widespread use of ineffective treatments, inefficient practices, and harm to patients[360,361]. Even though they are an essential resource, CPGs have historically exhibited low-quality reporting.[362] The ramifications of low reporting quality in CPGs are broad, but most pressing is the lack of a distinction between poor methods and poorly reported methods. In practice, the two may be indistinguishable. For example, if CPG developers perform a narrow, inadequate search of the literature, their subsequent recommendations may not be reproducible or trustworthy. Similarly, if the CPG developers do not report their search strategy, the question remains as to whether the recommendations are trustworthy. The quality of CPG reporting is as important as its methodological quality.

In oncology, new drug approvals may result in rapid changes to patient care. Articulating the available evidence, its strength, and its limitations to physicians is vital. The NCCN — arguably the premier guideline organization in the United States[363] — has a policy to update their CPGs "at least annually."[44] This policy of annual updates highlights the urgent need for clear reporting of current and future CPGs.

Two popular instruments exist for assessing the quality of CPGs in healthcare: The Reporting Items for practice Guidelines in HealThcare (RIGHT) statement[94] and the Appraisal of Guidelines for Research and Evaluation (AGREE) II instrument[95]. The AGREE-II instrument includes items related to the methodological (e.g., quality of search strategy, inclusion of stakeholder preferences) and reporting quality of CPGs, whereas the RIGHT statement focuses solely on reporting quality (e.g., providing a summary of recommendations, disclosure of funding source). Neither was created as a handbook for developing guidelines. According to the RIGHT Statement authors, the RIGHT Statement is not designed to assess the inherent quality of a guideline.[94] Rather, the RIGHT Statement is designed to complement tools that are designed to assess the inherent quality of a guideline, such as the AGREE-II instrument.

Given the comprehensiveness and importance of the NCCN CPGs to oncology practice[363], the aim of this investigation is to highlight the strengths and weaknesses in the reporting of NCCN guidelines. By doing so, we aim to improve the delivery of oncology evidence to oncologists and improve patient care. In this study we applied the RIGHT Statement and AGREE-II instrument to 49 NCCN guidelines for the treatment of cancer by cancer site.

### *Methods*

A version of this manuscript is available as a preprint via bioRxiv[364]. Since NCCN guidelines update frequently throughout a calendar year, we downloaded the PDF of all 49 NCCN treatment guidelines on March 21, 2018 from the NCCN website under the heading "NCCN Guidelines for the Treatment of Cancer by Site". To be included in this study, a guideline must have a written Discussion section, which is the equivalent to the guideline narrative. Prior to data extraction, CW, CC, and DT reviewed the RIGHT statement and AGREE-II instrument manuals to become familiar with the checklist items.[94,95] We met and devised a Google Form for both tools. CW, CC, and DT extracted data for all items from each tool independently, while masked to each other's decisions. Since the NCCN does not detail their full methods in each CPG, and provides a full explanation of many aspects of their methods on their website (www.nccn.org), we extracted data from the CPG and website policy documents. Any discrepancies in data extraction were resolved via consensus discussion. After extraction and validation of all Google Form responses, we exported these responses to a Google Sheet. We used this Google Sheet to calculate summary statistics. We correlated the RIGHT and AGREE-II scores using Stata 15.1 and the commands *pwcorr*, for a Pearson's r, and *graph twoway scatter* for a two-way scatter plot. Raw AGREE-II scores were used, rather than scaled scores, with a maximum value of 161 (23 items, 7-point Likert scale) indicating a judgement of perfect methodological quality across all domains for a CPG.

The design of the RIGHT Statement parallels other statements and reporting guidelines, such as CONSORT for clinical trials or PRISMA for SRs, and consists of a 35-item checklist and an Explanation and Elaboration document.[94]. For each of the items we assigned a numeric score of 1 (full adherence), 0.5 (partial adherence), or 0 (no adherence). An example of partial adherence may be if a guideline provides a partial explanation of cancer epidemiology, explaining only the prevalence and incidence of the disease. Full explanation includes a description of prevalence/incidence, morbidity, mortality, and burden (including financial). We present summary data using the described scoring convention for each of the 35-items. Rather than dichotomizing the data in an attempt to separate CPGs into high, medium, or low reporting quality groups, we present data as continuous and out of the maximum possible score of 35. This decision was made because there is no guidance for what constitutes high, medium, or low-quality reporting quality in CPGs.

The AGREE-II instrument is organized differently, and consists of 23 items divided into 6 domains, with each item scored on a 1 (strongly disagree) to 7 (strongly agree) Likert-type scale. In accordance with the AGREE-II manual,[95] we calculated a scaled domain score for each domain for each CPG. The scaled domain score is calculated as follows:

$$(Obtained\,Score\;-\;Minimum\,Score)\;\div\;(Maximum\,Score\;-\;Minimum\,Score)$$

The scaled domain score can be converted to an average rating (1 to 7 scale) by multiplying the scaled domain score by 7. The obtained score is calculated for each domain and is the sum of all rater scores in that domain. The minimum score is calculated by multiplying the minimum item score (1, strongly disagree), the number of raters (3, in this study), and the number of items in the domain. The maximum score is calculated similarly, but substitutes the maximum item score (7, strongly agree) for the minimum item score. Lastly, we made a consensus judgement about whether the CPG should be used in practice or not based on the 6 scaled domain scores for each CPG. We based our judgement of each NCCN CPG off the AGREE-II manual, which suggests answering whether a CPG should be used with "yes", "yes with modifications", or "no". We rendered our judgements by looking at the full scope of domain scores, rather than using dichotomous decision rules. The rationale for this decision was that each domain has been shown to independently associated with CPG quality[365]Our primary objective was to assess CPG scores on the RIGHT statement and AGREE-II instrument. Since all NCCN guidelines were published after the RIGHT statement and AGREE-II instrument were published, they are all eligible for analysis. As neither the RIGHT statement or AGREE-II instrument can judge the clinical usefulness of a guideline, our study is designed to only focus on the methodological and reporting quality of each guideline.

### ***Results***

We identified 49 NCCN CPGs for the treatment of cancer by site. The Uveal Melanoma CPG was excluded because the Discussion section (the narrative section of NCCN guidelines) was under development and not written. All of our data, including data for each individual item on the RIGHT statement and AGREE-II instrument, are publicly available via the Open Science Framework[366].

#### *RIGHT Statement*

The NCCN guidelines were largely homogenous, and many key methodological items were reported clearly in policy documents on the NCCN website. Table 13 shows each NCCN guideline and its adherence to all RIGHT statement items. Notable strengths of the NCCN CPGs were the reporting of conflicts of interest for all authors (items 19a and

19b), complete description of pertinent subgroups (item 7b), and the clarity of CPG recommendations (item 13a). Notable deficiencies were the description of stakeholder involvement (e.g., patient views and preferences) [item 14a], the cost and resource implications of therapies (item 14b), which outcomes were prioritized when formulating recommendations (item 10b), and the approach to assess the certainty of the quality of evidence (item 12).

### *AGREE-II*

Table 14 shows the scaled domain scores for each NCCN CPG. Using the AGREE-II instrument we were able to assess CPG scores in six domains, each essential to a methodologically robust CPG. No guideline scored extremely low for any domain. The fourth domain (Clarity of Presentation) and sixth domain (Editorial Independence) scored the highest, overall. The Clarity of Presentation domain asks whether the recommendations are specific and unambiguous, if alternative treatment options were mentioned, and if the key recommendations are easily identifiable. The sixth domain asks questions about the influence of the funding source on CPG development and whether conflicts of interest were disclosed. The lowest, individual domain score was 36.1% in the Applicability domain for the Acute Lymphoblastic Leukemia CPG. This score indicates that average score (1 to 7 scale) for this domain was approximately 2.5. With respect to overall domain scores across all guidelines, the Stakeholder Involvement domain scored the lowest with an average score of 48.6% (e.g., 3.4 out of 7). The Stakeholder Involvement domain asks questions related to the description of guideline development members, the incorporation of target population views and preferences, and the identification of target users of the guidelines.

### *Correlation of RIGHT and AGREE-II scores*

There was a small, negative correlation between RIGHT and AGREE-II scores (r = -.25). The negative correlation is likely driven by the 4 guidelines that adhered to only 19/35 (54.2%) of RIGHT items, while maintaining relatively high AGREE-II scores. Overall, most data clustered between RIGHT scores of 19.5 - 20.5 and AGREE-II scores of approximately 105-115. Visual inspection of our data shows that many CPGs had identical RIGHT scores, with slight variations in their AGREE-II scores.

| **Guideline** | **RIGHT Statement domain** | | | | | **Funding (4)** | **Total (35)** |
|---|---|---|---|---|---|---|---|
| | Basic Info (6) | Background (n = 8) | Evidence (n = 5) | Recommendations (n = 7) | Quality assurance (n = 2) | | |
| *Acute Lymphoblastic Leukemia* | 4.0 | 6.5 | 1.5 | 3.5 | 1.0 | 4.0 | 21 (60.0%) |
| *Acute Myeloid Leukemia* | 4.0 | 5.0 | 1.5 | 3.5 | 1.0 | 4.0 | 19.5 (55.7%) |
| *Amyloidosis* | 5.0 | 5.0 | 1.5 | 3.5 | 1.0 | 4.0 | 20.5 (58.6%) |
| *Anal* | 5.0 | 5.0 | 1.5 | 3.5 | 1.0 | 4.0 | 20.5 (58.6%) |
| *B-Cell* | 4.0 | 6.0 | 1.5 | 3.5 | 1.0 | 4.0 | 20.5 (58.6%) |
| *Basal Cell* | 4.0 | 5.0 | 1.5 | 3.5 | 1.0 | 4.0 | 19.5 (55.7%) |
| *Bladder* | 4.0 | 5.0 | 1.5 | 3.5 | 1.0 | 4.0 | 19.5 (55.7%) |
| *Bone* | 4.0 | 5.0 | 1.5 | 3.5 | 1.0 | 4.0 | 19.5 (55.7%) |
| *Breast* | 4.5 | 5.0 | 1.5 | 3.5 | 1.0 | 4.0 | 20 (57.1%) |
| *Cervical* | 4.0 | 4.5 | 1.5 | 3.5 | 1.0 | 4.0 | 19 (54.3%) |
| *Chronic Lymphocytic Leukemia* | 4.0 | 5.0 | 1.5 | 3.5 | 1.0 | 4.0 | 19.5 (55.7%) |
| *Chronic Myeloid Leukemia* | 4.0 | 5.0 | 1.5 | 3.5 | 1.0 | 4.0 | 19.5 (55.7%) |
| *Central Nervous System* | 5.0 | 5.0 | 1.5 | 3.5 | 1.0 | 4.0 | 20.5 (58.6%) |
| *Colon* | 5.0 | 5.0 | 1.5 | 3.5 | 1.0 | 4.0 | 20.5 (58.6%) |
| *Cutaneous B-Cell* | 4.0 | 3.5 | 1.5 | 3.5 | 1.0 | 4.0 | 18 (51.4%) |
| *Dermato-Protruberans* | 4.0 | 5.0 | 1.5 | 3.5 | 1.0 | 4.0 | 19.5 (55.7%) |
| *Esophageal* | 4.0 | 5.0 | 1.5 | 3.5 | 1.0 | 4.0 | 19.5 (55.7%) |
| *Gastric* | 5.0 | 5.0 | 1.5 | 3.5 | 1.0 | 4.0 | 20.5 (58.6%) |
| *Hairy Cell* | 5.0 | 5.0 | 1.5 | 3.5 | 1.0 | 4.0 | 20.5 (58.6%) |
| *Head/Neck* | 4.0 | 4.0 | 1.5 | 3.5 | 1.0 | 4.0 | 18.5 (52.9%) |
| *Hepatobiliary* | 4.0 | 5.0 | 1.5 | 3.5 | 1.0 | 4.0 | 19.5 (55.7%) |
| *Hodgkin* | 4.0 | 5.5 | 1.5 | 3.5 | 1.0 | 4.0 | 20 (57.1%) |
| *Kaposi* | 4.0 | 5.0 | 1.5 | 3.5 | 1.0 | 4.0 | 19.5 (55.7%) |
| *Kidney* | 4.0 | 5.0 | 1.5 | 3.5 | 1.0 | 4.0 | 19.5 (55.7%) |
| *Myelodysplastic Syndrome* | 4.5 | 5.0 | 1.5 | 3.5 | 1.0 | 4.0 | 20 (57.1%) |
| *Melanoma* | 5.0 | 5.0 | 1.5 | 3.5 | 1.0 | 4.0 | 20.5 (58.6%) |
| *Merkel* | 4.5 | 5.0 | 1.5 | 3.5 | 1.0 | 4.0 | 20 (57.1%) |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| *Mesothelioma* | 4.0 | 5.0 | 1.5 | 3.5 | 1.0 | 4.0 | 19.5 (55.7%) |
| *Myeloproliferative Neoplasms* | 4.0 | 5.0 | 1.5 | 3.5 | 1.0 | 4.0 | 19.5 (55.7%) |
| *Myeloma* | 4.0 | 5.0 | 1.5 | 3.5 | 1.0 | 4.0 | 19.5 (55.7%) |
| *Neuroendocrine* | 4.0 | 5.0 | 1.5 | 3.5 | 1.0 | 4.0 | 19.5 (55.7%) |
| *Non-small cell lung* | 4.0 | 5.0 | 1.5 | 3.5 | 1.0 | 4.0 | 19.5 (55.7%) |
| *Occult primary* | 4.0 | 5.5 | 1.5 | 3.5 | 1.0 | 4.0 | 20 (57.1%) |
| *Ovarian* | 4.0 | 5.0 | 1.5 | 3.5 | 1.0 | 4.0 | 19.5 (55.7%) |
| *Pancreatic* | 4.0 | 5.0 | 1.5 | 3.5 | 1.0 | 4.0 | 19.5 (55.7%) |
| *Penile* | 5.0 | 5.0 | 1.5 | 3.5 | 1.0 | 4.0 | 20.5 (58.6%) |
| *Prostate* | 4.0 | 6.0 | 1.5 | 3.5 | 1.0 | 4.0 | 20.5 (58.6%) |
| *Rectal* | 4.0 | 5.0 | 1.5 | 3.5 | 1.0 | 4.0 | 19.5 (55.7%) |
| *Sarcoma* | 4.0 | 5.0 | 1.5 | 3.5 | 1.0 | 4.0 | 19.5 (55.7%) |
| *Squamous cell* | 4.5 | 5.0 | 1.5 | 3.5 | 1.0 | 4.0 | 20 (57.1%) |
| *Small cell lung* | 4.0 | 4.0 | 1.5 | 3.5 | 1.0 | 4.0 | 18.5 (52.9%) |
| *T-Cell* | 5.0 | 6.0 | 1.5 | 3.5 | 1.0 | 4.0 | 21.5 (61.4%) |
| *Testicular* | 4.0 | 5.0 | 1.5 | 3.5 | 1.0 | 4.0 | 19.5 (55.7%) |
| *Thymus* | 5.0 | 5.0 | 1.5 | 3.5 | 1.0 | 4.0 | 20.5 (58.6%) |
| *Thyroid* | 4.0 | 5.0 | 1.5 | 3.5 | 1.0 | 4.0 | 19.5 (55.7%) |
| *Uterine* | 4.0 | 5.0 | 1.5 | 3.5 | 1.0 | 4.0 | 19.5 (55.7%) |
| *Vulvar* | 5.0 | 5.0 | 1.5 | 3.5 | 1.0 | 4.0 | 20.5 (58.6%) |
| *Waldenstrom* | 4.0 | 5.0 | 1.5 | 3.5 | 1.0 | 4.0 | 19.5 (55.7%) |

*Table 13. Adherence to RIGHT statement items overall and in each domain for all NCCN guidelines.*

Certain outliers are visible in the scatter plot, which have been labelled with the CPG name. Notable outliers are the guidelines for Merkel Cell Carcinoma and Primary Cutaneous B-Cell Lymphoma. The Merkel Cell Carcinoma guideline scored lowest on AGREE-II, but average on RIGHT. This guideline was judged to score relatively low on three methodological domains: Stakeholder Involvement, Rigor of Development, Applicability. None of these items had direct overlap with RIGHT statement items, so the Merkel Cell Carcinoma guideline was still capable of achieving an average score in terms of reporting quality. On the other hand, the Primary Cutaneous B-Cell Lymphoma guideline scored lowest on the RIGHT statement, but above average on AGREE-II. In absolute terms, the Primary Cutaneous B-Cell Lymphoma guideline only scored 2 items lower than most other guidelines.

| Guideline | Scope | Stakeholder Involvement | Rigor | Clarity | Applicability | Editorial Independence |
|---|---|---|---|---|---|---|
| *Acute Lymphoblastic Leukemia* | 77.8% | 42.6% | 61.8% | 85.2% | 36.1% | 94.4% |
| *Acute Myeloid Leukemia* | 79.6% | 38.9% | 59.0% | 87.0% | 40.3% | 94.4% |
| *Amyloidosis* | 42.6% | 50.0% | 57.6% | 74.1% | 48.6% | 94.4% |
| *Anal* | 72.2% | 37.0% | 54.9% | 81.5% | 40.3% | 94.4% |
| *B-Cell* | 74.1% | 50.0% | 56.9% | 81.5% | 52.8% | 94.4% |
| *Basal Cell* | 77.8% | 53.7% | 61.8% | 85.2% | 62.5% | 94.4% |
| *Bladder* | 74.1% | 42.6% | 57.6% | 90.7% | 51.4% | 94.4% |
| *Bone* | 77.8% | 51.9% | 67.4% | 85.2% | 59.7% | 94.4% |
| *Breast* | 79.6% | 50.0% | 70.8% | 79.6% | 62.5% | 94.4% |
| *Cervical* | 68.5% | 51.9% | 63.9% | 87.0% | 63.9% | 94.4% |
| *Chronic Lymphocytic Leukemia* | 83.3% | 51.9% | 57.6% | 83.3% | 65.3% | 94.4% |
| *Chronic Myeloid Leukemia* | 70.4% | 40.7% | 65.3% | 87.0% | 45.8% | 94.4% |
| *Central Nervous System* | 77.8% | 53.7% | 58.3% | 81.5% | 66.7% | 94.4% |
| *Colon* | 83.3% | 42.6% | 61.1% | 70.4% | 51.4% | 94.4% |
| *Cutaneous B-Cell* | 63.0% | 50.0% | 43.8% | 77.8% | 61.1% | 94.4% |
| *Dermato-Protruberans* | 77.8% | 50.0% | 65.3% | 88.9% | 63.9% | 94.4% |
| *Esophageal* | 81.5% | 50.0% | 66.0% | 87.0% | 63.9% | 94.4% |
| *Gastric* | 66.7% | 40.7% | 61.1% | 83.3% | 50.0% | 94.4% |
| *Hairy Cell* | 70.4% | 42.6% | 63.9% | 85.2% | 50.0% | 94.4% |
| *Head/Neck* | 74.1% | 51.9% | 65.3% | 81.5% | 65.3% | 94.4% |
| *Hepatobiliary* | 77.8% | 48.1% | 64.6% | 83.3% | 59.7% | 94.4% |
| *Hodgkin* | 70.4% | 42.6% | 68.1% | 87.0% | 52.8% | 94.4% |
| *Kaposi* | 70.4% | 48.1% | 63.9% | 83.3% | 54.2% | 94.4% |
| *Kidney* | 74.1% | 53.7% | 61.8% | 85.2% | 65.3% | 94.4% |
| *Myelodysplastic Syndrome* | 63.0% | 40.7% | 43.8% | 68.5% | 40.3% | 94.4% |
| *Melanoma* | 72.2% | 51.9% | 62.5% | 81.5% | 61.1% | 94.4% |
| *Merkel* | 68.5% | 48.1% | 65.3% | 85.2% | 61.1% | 94.4% |
| *Mesothelioma* | 81.5% | 40.7% | 62.5% | 87.0% | 51.4% | 94.4% |
| *Myeloproliferative Neoplasms* | 79.6% | 53.7% | 59.0% | 85.2% | 55.6% | 94.4% |
| *Myeloma* | 70.4% | 53.7% | 66.7% | 87.0% | 63.9% | 94.4% |
| *Neuroendocrine* | 74.1% | 46.3% | 67.4% | 87.0% | 69.4% | 94.4% |
| *Non-small cell lung* | 77.8% | 51.9% | 66.7% | 87.0% | 63.9% | 94.4% |

| | | | | | | |
|---|---|---|---|---|---|---|
| *Occult primary* | 74.1% | 51.9% | 66.0% | 87.0% | 63.9% | 94.4% |
| *Ovarian* | 85.2% | 51.9% | 62.5% | 87.0% | 61.1% | 94.4% |
| *Pancreatic* | 74.1% | 51.9% | 66.7% | 85.2% | 65.3% | 94.4% |
| *Penile* | 72.2% | 53.7% | 68.1% | 87.0% | 63.9% | 94.4% |
| *Prostate* | 81.5% | 46.3% | 67.4% | 85.2% | 56.9% | 94.4% |
| *Rectal* | 70.4% | 53.7% | 63.9% | 87.0% | 66.7% | 94.4% |
| *Sarcoma* | 75.9% | 53.7% | 61.8% | 87.0% | 58.3% | 94.4% |
| *Squamous cell* | 74.1% | 53.7% | 66.7% | 85.2% | 62.5% | 94.4% |
| *Small cell lung* | 70.4% | 42.6% | 50.7% | 85.2% | 48.6% | 94.4% |
| *T-Cell* | 75.9% | 44.4% | 62.5% | 85.2% | 50.0% | 94.4% |
| *Testicular* | 74.1% | 53.7% | 64.6% | 87.0% | 69.4% | 94.4% |
| *Thymus* | 77.8% | 46.3% | 64.6% | 87.0% | 52.8% | 94.4% |
| *Thyroid* | 63.0% | 53.7% | 64.6% | 87.0% | 68.1% | 94.4% |
| *Uterine* | 74.1% | 53.7% | 64.6% | 87.0% | 61.1% | 94.4% |
| *Vulvar* | 79.6% | 42.6% | 63.9% | 85.2% | 51.4% | 94.4% |
| *Waldenstrom* | 74.1% | 55.6% | 66.0% | 87.0% | 61.1% | 94.4% |
| **Average Scaled Domain Score** | **73.9%** | **48.6%** | **62.4%** | **84.4%** | **57.5%** | **94.4%** |

*Table 14. AGREE-II scores, using scaled percent adherence, in all domains and overall across all guidelines.*

### *Discussion*

In this investigation we found that NCCN CPGs demonstrate key strengths and weaknesses with respect to the reporting of key items essential to CPGs. For example, the NCCN CPGs require conflicts of interest disclosure, clearly describe all pertinent subgroups, and delineate key recommendations. On the other hand, the NCCN CPGs did not consistently describe how patient values and preferences were incorporated into recommendations, the financial burden of the recommendations, or describe the approach used to assess the certainty of the evidence underpinning the recommendations. The NCCN guidelines were incredibly uniform in how they are reported and conducted, which resulted in similar (or identical, in the case of the RIGHT statement) scores for most CPGs. This uniformity is reflected in the scatter plot. Across all NCCN guidelines, certain items, such as providing a summary of recommendations, were always reported. On the other hand, some items, such as describing the approach to assessing the certainty of the evidence, were never reported. The slight variations in AGREE-II scores for identical RIGHT scores is a product of 1-7 Likert scale format, which allows more variation in judgements than the RIGHT statement scoring system of full, partial, or no adherence. In light of the uniformity of our data, our findings should be interpreted to mean that there are significant shortcomings in the reporting and development of NCCN guidelines, but all of these shortcomings could be addressed at once by updating the central NCCN policies and procedures.

Nonetheless, compared to other CPGs scored with the AGREE-II instrument, those published by the NCCN appear to have as good or stronger methodological quality[365,367–369]. A recent evaluation in *JAMA Internal Medicine* of CPGs for the pharmacologic management of noncommunicable diseases in primary care found that three CPG characteristics are associated with high quality CPGs: greater than 20 authors, development at a government institution, and reported funding[370]. The NCCN is a non-profit organization and their CPGs are developed by a team of volunteers from member institutions and no external funding is received to develop the CPGs. All guidelines all have greater than 20 authors. So, the findings of this recent evaluation in *JAMA Internal Medicine* seem in line with our findings that NCCN CPGs are of comparable or higher methodological quality than other biomedical CPGs. However, the reporting quality of biomedical CPGs has been evaluated far less, owing to the fact that the RIGHT statement is the only available tool and was published in 2017. Only one study was identified which used the RIGHT Statement[371]. This lone study evaluated 539 CPGs in traditional Chinese medicine, finding that 17 of 35 (48.6%) RIGHT Statement were reported less than 10% of the time. In comparison, our study found that only 9 items were never fully reported. In an effort to provide the highest-quality recommendations to physicians for the treatment of different cancers, we encourage continued improvements to the NCCN guidelines. The AGREE-II instrument[95] was developed to assess CPG quality in six, equally essential domains ranging from describing the purpose of the CPG to the applicability of the CPG recommendations. We found that they scored well enough to continue being recommended in clinical practice, but key methodological items were not reported, thus highlighting areas where the delivery of oncology evidence can be improved. Since we assigned summary judgements related to the recommended use of NCCN CPGs in clinical practice in a continuous manner, each judgement of “Yes, with modifications” should be interpreted continuously. Since no two CPGs were scored identically for all 6 domains, each judgment of “yes, with modifications” should signal different improvements are needed in different orders of magnitude. Through applying the RIGHT Statement, which was created to be used alongside the AGREE-II instrument, we confirmed that improvements in the reporting of several key items would strengthen the impact of NCCN CPGs by increasing the clarity and comprehensiveness of the recommendations.

None of the NCCN CPGs described the process by which patient values and preferences were solicited and incorporated into the guideline recommendations, nor do they adhere to an accepted framework for grading the quality of evidence. The primary reason for incorporating patient values and preferences into CPG recommendations is that recommendations that are aligned with patient values may be more easily adopted and implemented[372–374]. Until recently, there were no firmly established processes for including patient values and preferences in CPG recommendations. To address this gap, the GRADE (Grading of Recommendations Assessment, Development and Evaluation) working group created the GRADE Evidence-to-Decision (EtD) framework[373]. Previously, the GRADE approach has been used to assess the quality and certainty of evidence underpinning CPG recommendations. The NCCN CPGs do not currently use the GRADE approach, or any similar framework, rather they seem to rely on guideline development member assessments of the quality of evidence. The NCCN members assess the quality of evidence over certain domains, but in an effort to improve the objectivity, applicability, and comparability of NCCN recommendations, we recommend adopting the GRADE approach. Concurrent adoption of the GRADE EtD framework would ensure the incorporation of patient values and preferences in all recommendations.

Additional, minor adjustments to the reporting of NCCN CPGs would improve the delivery of oncology evidence. First, stating key research questions that formed the basis for treatment recommendations in PICO (Population, Intervention, Comparator, Outcome) format would guide physicians through the purpose and scope of the guideline[80,375,376]. Due to how comprehensive the NCCN CPGs are, it may be that listing all PICO-format questions is not practical. Should this be the case, we recommend including a section in the CPG that clearly describes the scope, limitations, and gaps in the NCCN recommendations. A second, related adjustment includes listing the outcomes that were most important when developing the CPG recommendations. For example, if efficacy outcomes are the primary basis for the recommendations, or recommending one treatment over another, physicians would benefit from that understanding.

This study has key strengths and limitations. With respect to strengths, we used two formally published and peer-reviewed tools to assess the quality of reporting and methodological rigor of NCCN guidelines. We further used 3 data extractors to mitigate bias in our data analysis. Each author underwent identical, comprehensive training to ensure competency prior to data extraction. With respect to limitations, our assessment of methodological quality may be limited by a lack of reporting. In other words, simply because someone was not reported as having been done, does not mean it was not done. For example, it is possible that the views of patients were sought in the formulation of the guidelines, but if this was not reported or described, we were forced to assign a low score this AGREE-II item low.

In conclusion, we simultaneously recommend the continued use of NCCN CPGs to guide oncologists in patient care and efforts to improve the weaknesses we identified in this study. Each guideline contained strengths and weaknesses, and improving the weaknesses will enhance the applicability and comparability of the recommendations. We have outlined key recommendations that would improve the completeness of reporting and increase transparency. These recommendations include the adoption of the GRADE and GRADE-EtD approach, describing key questions in PICO format, and sorting which outcomes were important when developing recommendations. We believe that adopting these recommendations will not only improve the NCCN CPGs, but oncology clinical care as well.

# CHAPTER XIII

# DISCUSSION AND CONCLUSION

The aforementioned 10 studies investigated various forms of bias, transparency, and reproducibility in oncology clinical trials, SRs, and CPGs. All studies involved top-ranked journals, prestigious CPGs, or highly-regarded studies. Key results from each study are summarized in Table 15. In-depth discussions and conclusions were included in prior chapters, which leaves room for a discussion of how all these studies fit together and how the results may be interpreted together to improve the rigor and reproducibility of cancer medicine evidence. In particular, the various forms of bias and irreproducibility that have been discussed likely contribute to hype, financial toxicity, and other practices that do not align with what patients may expect from their oncologist. Namely, patients likely expect impartiality and strong critical appraisal from their oncologist and from the oncologists who recommend medications. Based on the results presented above, it is likely that an oncologist's ability to make evidence-based decisions is more difficult in light of various forms of bias, lack of transparency, and lack of understanding inherent to the design, conduct, and reporting of oncology research. In other words, the state of oncology research is such that if one wants to offer strong critical appraisal, they will likely meet roadblocks — known and unknown — that make their pursuit more difficult.

First, it is prudent to discuss clinical trials, since trials are the most important primary study in oncology. Like all medical research, oncology research contains barriers to fully trustworthy, translatable results. Such barriers may be likened to "random error", while other barriers, based on the results presented in this dissertation, are more systematic. Namely, the high prevalence of conflicts of interest with the pharmaceutical industry is more likely to be a systematic error in oncology research than a random one. Chapter V shows that approximately 75% of included oncologist-authors had financial relationships with industry and that approximately 40% of those authors did not disclose all of their conflicts of interest. The pattern that emerged from that data is one of inattention to detail, rather than overt deception. After all, these authors are likely aware of the Open Payments database , so if these authors were trying to hide their financial relationships, they would be unable to do that job well.

| Chapter & Brief Title | Key results |
|---|---|
| III. Reporting guidelines in oncology journals | 16/21 journals use CONSORT; 16/21 journals mention trial registration |
| | Journal mention of CONSORT increased use of CONSORT, although the effect was imprecise |
| | Mentioning trial registration did not seem to affect trial registration rates |
| IV. Reporting of oncology trial interventions | Reporting of trials leading to FDA-approval is largely homogeneous, with key deficits |
| | Targeted focus on the details of trial centers, intervention administrators, and compliance assessments would improve the reporting and translation of trial findings |
| V. Financial conflicts of interest in trials | Oncology trial authors often have high-dollar, undisclosed conflicts of interest |
| | NEJM, Lancet Oncology, and Lancet Haematology had the highest proportion of authors with undisclosed conflicts of interest |
| VI. Spin in trials | 46/124 clinical trials had spin in the abstract, most often in the conclusions by ignoring overall survival data |
| | There was no difference in the rate of spin when overall survival was a primary or secondary endpoint, indicating that spin is used for different reasons across trials |
| VII. Interim analyses | Interim analyses with surrogate endpoints are overwhelmingly statistically significant, while final results with overall survival are more often nonsignificant |
| | Interim analyses generate more hype and attention than final analyses |
| | Many interim analyses use surrogate endpoints that are not valid predictors of survival |
| VIII. Selective outcome reporting in advertisements | Advertisements often exclude unfavorable secondary outcomes, while including favorable ones |
| | Advertisements often air or print before final overall survival data is published |
| | Advertisements may contribute to hype and biased prescribing of drugs based on preliminary data alone |
| IX. Quality of noninferiority trials | Poor design and methods were common in included trials |
| | Notably, confidence intervals and P-values were not consistent, justification for the margin was absent, and per-protocol analyses were omitted |
| X. Reproducibility of oncology meta-analyses | Only approximately 65% of meta-analyses were reproducible at face value |
| | The majority of non-reproducible meta-analyses were presented in table form, rather than figures |
| | Key omissions included how missing and unpublished data were handled, what summary statistic was used, |
| XI. Risk of bias and reporting quality in systematic reviews | Only 3 colorectal systematic reviews were at low risk of bias |
| | Bias was most common in protocols, eligibility, and study identification domains |
| | Overall, adherence to PRISMA was good, with no difference between reviews that mentioned and did not mention PRISMA, indicating well-defined standards for reporting throughout the systematic review community |
| XII. Reporting and quality of practice guidelines | NCCN guidelines have strong guidance for conflicts of interest, clearly describe recommendations, and clearly define important subgroups |
| | These guidelines lack involvement of patient stakeholders, consideration of cost, and a clear description of how uncertainty was adjudicated in the evidence base |

*Table 15. Key findings from each of the included studies for this dissertation.*

Trying to explain the inattention in disclosing conflicts of interest is difficult, but it is likely that journal response to undisclosed conflicts of interest contributes to author inattention. Despite journals having access to the same data we used to check the disclosure accuracy, Chapter V shows that authors consistently did not disclose all conflicts, and complete accuracy was not verified. No public corrections had been made to disclosures at the time of our study. Many even believe that close working relationships with industry are advantageous for patients and physicians alike[235]. Drug firms have a near-monopoly on clinical trials now because of the immense global cost to bring a drug to market. Oncologists who publish these trials often see an increase in the amount of money they receive as a result of their financial relationships, per our preliminary findings in Chapter V. This system creates numerous incentives for key opinion leaders in oncology to praise any positive finding, because one's career and financial status advances as more trials are published.

For oncologists who see patients, rather than only conducting clinical trials, there is ample evidence that financial relationships with industry affect physician behavior. The strongest evidence worth discussing is from the National Bureau of Economic Research, whose 2020 publication using Medicare Part D data showed that physicians who receive money from industry for a drug are more likely, for a time after receiving money, to prescribe that drug compared to unpaid physicians[154]. The efficacy of the drug for which payment was received did not explain the differential prescription. Over time, if paid physicians remained unpaid, their prescribing habits fell back in line with physicians who had been unpaid all along. Other relational and prescribing habits were also observed that could add to the deleterious effects on patients. The drug firms were found to change their advertising and payment campaigns to physicians when generic drug competition began, by advertising new formulations, such as an extended-release version of the branded drug. Altogether, payments from industry, even in the form of food or drug samples, are likely to benefit oncologists in clinical practice just like they do clinical trialists. In addition, these payments to physicians are not benign for patients and likely contribute to financial toxicity and strain on the healthcare economy.

Financial conflicts of interest do not only affect the interpretation, critical appraisal, and translation of clinical trials to patient care: they also affect CPGs. A seminal paper in oncology research was published in 2016 and described the extensive conflicts of interest held by oncologists-authors of NCCN guidelines. These guidelines, as previously mentioned, are the most widely recognized and influential guidelines in the United States. These guidelines made zero mention of patient values and preferences when formulating guideline recommendations which is problematic for several reasons. First, at face value, this strategy is unlikely to result in recommendations that align with what patients want. Second, the omission of patient values and preferences means that the only potential external influences on oncologist-authors are ones from their social and academic circles, which is composed of drug firm sponsors and fellow authors who likely have strong ties to drug firms. Last, patients with cancer have to navigate financial[6], physical[377], social[378], and psychological harms[379] in the face of a potentially terminal disease. With rapid development of novel cancer therapeutics, the toxicity profile of standard cancer therapy has dramatically changed and the messaging around these changes is that the new drugs are less toxic when in fact the toxicities are just different. Older cancer therapies were administered in cycles, with off weeks scheduled in advance where patients would not receive any medications. The weeks with medication were difficult, sometimes requiring proactive hospitalization to combat expected harms[380]. Older chemotherapy regimens were more likely to cause higher-grade toxicities[381], but were given in cycles, rather than continuously. Novel cancer treatments have lesser-grade toxicity profiles, but are sometimes given daily, which raises the question of which is preferable: higher-grade toxicity less often or lower-grade toxicities daily? The decision and cost considerations may be different for individuals, but no such preferences were considered in NCCN guidelines. This omission appears ominous when one considers that the new medications that are often favored in the NCCN guidelines are ones for which oncologist-authors have received extensive payments from drug firms and cost significantly more for patients and the healthcare economy.

The NCCN guidelines are unique in that they are one of 5 guidelines that comprise the Centers for Medicare and Medicaid services (CMS) compendium, which means that anything recommended in the NCCN guidelines — FDA-approved or not — must be reimbursed by CMS[89]. This creates a strong incentive for a company’s drug to be included in the NCCN guidelines. Given what is known about financial relationships with industry and changes in prescribing practices, it is no wonder that a recent analysis of NCCN guidelines found that NCCN guidelines recommended a cohort of 47 novel drugs for 113 indications when only 69 indications were approved by the FDA[48]. Such "off-label" recommendations give patients and oncologists more treatment options, which is a strategy lauded by the FDA[3], but may decrease the value of each drug on average. To understand this over-recommendation and why it is unlikely to help patients, it is important to understand how oncology trials are designed. It is common for an oncology trial to have 2 primary endpoints: a surrogate endpoint and clinical endpoint. The trial is designed to accrue patients and outcome events until both endpoints are fully powered, but the surrogate endpoint accrues events faster, resulting in data that is mature more quickly. These mature data are then published and generate much hype, attention, and acclaim for authors. This phenomenon is described in Chapter VII. The reason this strategy took hold in oncology research is because of the long-recognized problem that patients are dying from terminal cancer every day, and that clinical trials take years. This traditional delay from preclinical tests to three phases of clinical trials to market authorization is designed by the FDA to protect patients from unsafe or ineffective drugs. This delayed regulatory pathway came under scrutiny during the AIDS epidemic in the 1990s when patients were dying rapidly, with promising drugs available, but not studied to the satisfaction of the regulatory agencies[382]. The solution was to create a novel regulatory pathway — Accelerated Approval — where drugs that were reasonably likely to predict clinical benefit, based on data from surrogate endpoints, could be approved for market use[383]. The caveat is that these drugs would have to show clinical benefit in a follow up trial within 3 years. A recent analysis of oncology drugs showed that early publications with surrogate endpoint data are only 11 months faster on average than trials that waited for OS data[57]. Those 11 months are meaningful, but a far-cry from the multi-year estimates many advocates of surrogate endpoints tout. This 11-month estimate is low enough that it raises questions about whether an unintentional loophole was created by the FDA. After accelerated approval, the 3-year countdown for drug firms to submit confirmatory data with OS to the FDA starts. If the average trial is only delayed by OS by 11 months, that means some drugs may remain on the market for 2 additional years before submitting OS data, which if negative may result in removal of market authorization. Constant and careful review of FDA drug policy is necessary to ensure that the system works in the best interests of patients, rather than drug firms.

It is plainly clear that surrogate endpoints are often poor predictors of patient-centered endpoints. This data is based on multiple SRs and meta-analyses that studied the correlation of surrogate endpoint with OS and quality of life results[61,62,64]. Despite that, surrogate endpoints are used by the FDA and relied upon by NCCN guideline authors to approve or recommend novel drugs for patient care. As shown in Chapter VII the oncologists hype early results with only surrogate endpoint data, as it may be a widely held belief that surrogate endpoints are truly meaningful. It is possible that these surrogate endpoints are favored by the drug firms as well. After all, it is known that surrogate endpoints do not predict clinical benefit, but positive surrogate endpoint results is the key to market approval and recommendation by the FDA and NCCN, which can result in billion-dollar profits in a matter of years[384]. Drugs are rarely removed from the market for failing to demonstrate clinical benefit after accelerated approval, but even if they were, the three-year window in which drug firms must present confirmatory data to the FDA is enough time to make a profit on the drug. As a result, there is no incentive for for-profit drug firms to avoid surrogate endpoints when designing their trials. This system also makes it difficult for everyday oncologists to be skeptical in their clinical practice about new treatments. In a hypothetical scenario where an oncologist chose to not prescribe a novel therapy to a patient who dies because of his or her belief that surrogate endpoint data is not sufficient, the malpractice lawsuit would focus on the fact that the experts in the field — authors of NCCN guidelines — recommended the drug. The social pressure to conform is likely immense, and surrogate endpoints present one of the most challenging problems in oncology research

In addition to the problems that surrogate endpoints present to physicians in terms of being a barrier to independent critical appraisal and pursuits of maximal clinical benefit, surrogate endpoints present a problem for patients as well. It is unlikely that patients fully understand the difference between a surrogate endpoint and a clinical endpoint. This belief is based on a recent SR of patient values and preferences in regards to surrogate endpoints in cancer drug data[58] which found inconclusive results, due to inaccuracies in how the authors of included studies described surrogate endpoints to patients. If the authors were not able to define a surrogate endpoint to patients, it is unlikely that patients could define these endpoints themselves. This represents a barrier in oncology research to patient health literacy and self-efficacy. Some have called for a renaming of some surrogate endpoints, like PFS[385] which is a composite endpoint of time to tumor growth by 20%, development of new tumor lesions, or death — whichever occurs first[114]. Unsurprisingly, the former two items occur more quickly than the latter one, which means that PFS is hardly a measure of survival at all. In the aforementioned SR of 15 investigations of patient values and preferences toward surrogate endpoints, oncologists omitted the following items to patients: 1) that PFS may not predict OS (10/15), 2) that progression may not affect how you feel (10/15), and 3) that progression does not mean treatment is needed (14/15). Data from patients in these studies is likely wasted because it fails to capture opinions that are meaningful, since the opinions were based on a faulty definition of the endpoint.

Beyond clinical trials in journals and guidelines, oncology drug advertisements are a significant barrier to proper critical appraisal for patients and physicians. In Chapter VIII we describe how oncology advertisements selectively reported their efficacy endpoints based on statistical significance. The implication is that mostly positive results are being conveyed to patients and physicians in advertisements. Our findings fit within what is known about potential harms of drug advertisements, which include potentially misleading patients,[282,283] exaggeration of drug benefits,[284,285] omissions regarding quality of life,[286] and increased healthcare spending[279,280]. Empirical data for how advertisements affects physicians is unclear, but from what is known about how interactions with industry affect physician behavior, it is within reason to believe that misleading advertisements will similarly affect physician behavior. The strategy for these advertisements appears to be to fully accrue patient data for a surrogate endpoint, submit for FDA Accelerated Approval, receive said approval, and market the drug based on the preliminary data which is overwhelmingly favorable. A significant proportion of advertisements reviewed did not mention OS or were made public while OS data was accruing. As mentioned above, patients may not know the difference between a surrogate endpoint and a clinical endpoint, while doctors may be swayed because these advertisements are a form of industry interaction. The selective outcome reporting evident in these advertisements follows what has been shown in previous work on breast cancer-related toxicities[185] and oncology trials as a whole[189]. Selective outcome reporting is one form of misinterpretation of research findings that is endemic to medical research and oncology research alike.

There are other ways in which authors of oncology trials spin their research findings to misrepresent what was found. We studied this in a cohort of 124 oncology clinical trials that measured both a surrogate endpoint and OS. This study confirmed that oncology trial authors misrepresent their research findings, even in major oncology journals that exert a strong influence on clinical practice. How spin was used by authors in the included studies indicates one worrisome conclusion: oncology trial authors emphasize OS or surrogate endpoints – whichever is most favorable. The evidence for this is that when OS was a primary endpoint, spin was more common when the data were unfavorable. On the other hand, when a surrogate endpoint was the primary endpoint, spin was almost always used to ignore fully-mature, nonsignificant OS data. These oncology trials enroll patients until the surrogate endpoint and OS are fully powered, so the use of "primary" and "secondary" endpoint is different than in other forms of research where only the primary endpoint is the basis for the sample size calculation. In reality, these trials have 2 primary endpoints, one of which they call secondary, and when that "secondary" endpoint is not favorable, they dismiss that data in the abstract. This would not be impactful if the "secondary endpoint" that they were dismissing was a surrogate endpoint or something else that is of little clinical importance to patients. Given that there are 2 endpoints in oncology that truly matter to patients — OS and quality of life — it is worrisome that a pattern has emerged in oncology research where publication in a top-tier journal is possible without patient-centered data. This is akin to a cardiology trial dismissing unfavorable results for major adverse cardiac events to highlight that symptomatic angina was improved. Angina is important, like delayed tumor growth is important, however if patients were queried, one will expect that they care more about living longer and living better, which is best accomplished by reducing risk of heart attack, stroke, and death[386]. Spin, if it continues to pervade the oncology literature, will not only affect perceptions of oncology drug efficacy[52], but also continue to indicate that there is no clear mechanism to curtail the number of drugs that are approved for market use off surrogate endpoints that fail to demonstrate commensurate benefits in patient-centered endpoints[203]. Such a bias in oncology research will have downstream effects on patient care.

Biases previously discussed have been long-standing issues in oncology research. The increasing popularity of noninferiority trials presents new issues for oncology research. Based on our analysis, there is a lack of understanding of what constitutes a rigorous noninferiority clinical trial. In many ways, traditional knowledge about clinical trials is flipped for noninferiority trials. In particular, the intention-to-treat analysis may actually introduce bias, rather than prevent it[387]. One-sided *P* values may be used to determine statistical significance, which may create asymmetry if the width of the two-sided confidence intervals used for hazard ratios is not carefully reviewed. However, the design item with the most potential for harm is the noninferiority margin. These margins are meant to preserve a portion of the active control effect, such that the new treatment is still moderately effective while offering other benefits, like lower cost or toxicity[388]. Concerns about inappropriate noninferiority margins were discussed in the approval of lenvatinib, whose margin at face value may have seemed appropriate if one did not critically appraise the trial on which the margin was based[311]. Such an example, discussed at length in Chapter IX, highlights the fact that the effect of bias on the practice of medicine can multiply as future research is conducted.

Meta-analyses have been the prototypical example of how bias and lack of rigor can ripple through multiple generations of research studies. The catch-phrase for how meta-analyses can multiply bias is “garbage in, garbage out”[389]. It is possible that even if risk of bias evaluations are conducted for primary studies in SRs, that primary studies could be biased. For example, risk of bias evaluations may fail to capture the problems that plague sorafenib for the treatment of hepatocellular carcinoma[312], such as the fact that the trial’s inclusion criteria did not match real-world patients who are older, more sick, and less physically capable[312]. Additional concerns arise from the SR itself, rather than the included trials. In Chapter XI we describe how the SRs cited by NCCN guidelines were explicitly at high risk of bias or under suspicion for bias due to lack of clear reporting of methodological items. Specifically, basic principles for good science were not followed like publication of a protocol and broad database searches. The latter omission means that the sample of studies included in the SR may be biased toward a more favorable effect, given what is known about published studies and statistical significance[78]. The former omission is a barrier to critical appraisal and the reproducibility of oncology SRs. The inability for independent oncologists to verify the rigor of SRs is worrisome, and this worry is compounded by the fact that all the SRs we evaluated were cited by the NCCN guidelines. Recall, the NCCN guidelines are the premier set of oncology guidelines in the United States and are one of the five compendium guidelines for the Centers for Medicare and Medicaid services. Irreproducible SRs force implicit trust of the findings presented and from what is known, these findings are potentially skewed.

To reproduce an SR is theoretically much easier than a clinical trial. To fully reproduce a clinical trial, one would have to re-enroll a new set of patients and recreate the same conditions as the original trial. For an SR, one would only have to independently search for studies, extract data, and synthesize results. Some of these steps may be skipped if the search strategy is robust at face value and data are fully available. Reproducing the synthesis of findings, called computational reproducibility[390], requires at minimum the event rates and sample size to be presented in a meta-analysis. By these numbers, one can calculate effect sizes, then reproduce the model characteristics specified by the study authors. We found in Chapter X that even this low-level version of reproducibility was not feasible in a significant portion of meta-analyses included in SRs cited by NCCN guideline authors. Data were made publicly available once, and protocols were rare, confirming what was found in Chapter XI. Moreover, key methodological items that would be necessary knowledge if one were to go through the process of reproducing database searches and data extraction were missing. In other words, while words were written to describe the SR and meta-analytic methods, these studies were really a black box, where the process of how the final results were calculated is unknown. If data were available, this would not be a concern because availability of data is a crucial step to allow the inference of how data extraction was conducted. However, as previously stated, only 1 of 154 SRs cited by the NCCN guidelines made their data publicly available.

What can be done in the future to improve the rigor and reproducibility of oncology research? Based on the findings presented in this dissertation, advancements in 5 areas are proposed.

First, the easiest way to start re-structuring oncology research to work in the best interest of patients is to commit to improved reporting practices and transparency in clinical trials, SRs, and CPGs. Chapters III and IV describe how there are gaps in journal policies for submitting authors to follow reporting guidelines. CONSORT guidelines, for clinical trials, have been shown to increase the completeness of clinical trial reporting in major medical journals,[391] and many aspects of CONSORT have now woven themselves into medical research practices. For example, the simple addition of "randomized controlled trial" to the title of papers has improved the indexing of clinical trials, which facilitates more robust SRs and meta-analyses[392]. Other CONSORT items, like publication of a protocol and statistical analysis plan are commonplace now. CONSORT was so successful in highlighting good and bad reporting practices that researchers eventually developed the TIDieR checklist to narrow the reporting focus on trial interventions[206]. Our analysis of oncology trials using the TIDieR checklist found that the major important components of interventions were described well, but items that would facilitate translation to all types of oncology practices were absent. Evidence suggests a difference in patient outcomes between academic centers where clinical trials are conducted[216], and community oncology clinics, where the majority of oncology care takes place[217,218]. Small shifts in how protocols are structured to better report trial methods would allow oncologists and other researchers an insight into what differences may be contributing to different outcomes for patients treated outside a clinical trial. Altogether, none of the aforementioned improvements in reporting will be as influential in shifting the landscape of oncology research as advancements in open data. The open science movement has advocated for publicly-available, open data to improve the reproducibility and translation of clinical evidence to practice[132]. Much of the trial data in oncology is proprietary and owned by pharmaceutical companies who sponsor trials. While these companies offer access to data through an application, significant barriers exist to being approved and using the data[393]. The goal of open data is not to simply hunt down errors that led to false results, but also to allow for better meta-analyses based on individual patient data that are tailored to specific populations that one sees on a daily basis. If a global trial is majority white individuals from the USA and UK, an Indian oncologist may wish to exclude these patients to determine the effect size on those individuals in his/her country. Without open data, there will continue to be a reliance on what is believed to be true, rather than a more precise estimate based on data with a narrow focus.

Second, oncology research must expand its focus on patient values. A recent analysis of the mean survival gain of oncology interventions found a 2.1-month average increase[116]. In such cases where the OS benefit is low or zero, quality of life becomes the default patient important endpoint to which trial sponsors and regulatory agencies can turn to determine if a drug improves patient outcomes. Multiple studies suggest that the prevalence of favorable quality of life results in cancer clinical trials ranges from 40-50%[394,395]. Unfortunately, a recent retrospective study of cancer drugs approved by the European Medicines Agency found that only 7/68 (10.3%) approved drugs between 2009 and 2013 had favorable quality of life data at the time of market approval[395]. Even worse, an analysis of quality of life assessments found that quality of life is often only measured during the time the intervention is given in the clinical trial setting, rather than until patient death[396]. This same study showed that most studies reporting quality of life until death showed that the intervention had worse quality of life than the control. It may be that the details of how long we measure quality of life can be set aside for now, since the logical first step is increasing oncologist reliance on quality of life as an outcome of interest for patient care. Our analysis of the reporting of NCCN guidelines showed that no explicit inclusion of patient-values and preferences was identified. A recent SR found that oncologist consideration of patient values and preferences was a facilitator of shared decision making, while poor physician communication was a barrier[397]. Given the NCCN guidelines' reputation for guiding the oncology community, it is recommended that leadership in patient-centered care begin at the NCCN guideline level. These guidelines may begin by consulting patients, patient advocates, and psycho-oncology specialists to determine whether recommended treatments align with patient values related to efficacy, cost, and toxicity.

Third, it is important that oncologists are clear and precise when they discuss drug efficacy. Surrogate endpoints measure drug activity, such as whether a drug can temporarily slow tumor growth[114]. These surrogate endpoints do not measure survival, and their inability to reliably predict survival demonstrates how many factors — cancer-related and not — contribute to a patient's survival. When oncologists recommend a drug in a guideline based on surrogate endpoint data, there needs to be an understanding that this data is not final, and that there is likely a significant degree of imprecision to be expected with respect to patient outcomes. No such nuance is conveyed by key opinion leaders in oncology, as evidenced by the degree of hype and attention that surrogate endpoint results obtain versus the subsequent OS (Chapter VII). There is no strong evidence that patients understand the difference between PFS and OS[58]. The over-reliance on surrogate endpoints is not solely the fault of oncologists themselves. The drug firms who sponsor clinical trials make full use of the FDA's accelerated approval system. The majority of new drugs receive initial approval from a surrogate endpoint, and this proportion is increasing[3]. In the first three years of market approval, many of these drugs will garner million, even billion, dollar profits for the drug firms[398]. In the worst-case scenario, where the FDA revokes market authorization based on insufficient confirmatory data, drug firms who obtain accelerated approval will still be profitable. The system, which began as a way to give patients access to novel drugs for a disease with no available treatments, is now being used to give patients additional treatment options that are costlier than ever, with no known benefit on survival or quality of life.

Fourth, the oncology profession must disincentivize financial relationships with drug firms and reward divestiture. The evidence that FCOI acutely affects physicians prescribing habits is overwhelming[154,223,399]. These habits shift toward more costly drugs with unclear advantages. For some medical treatments, these shifts are less meaningful, and may not result in much harm to patients or strain on the healthcare economy. For example, there is flexibility in which long-acting inhalers to prescribe to a patient, and costs are not much different[400]. Cancer drugs may vary in price by tens of thousands of dollars and may have very different toxicity profiles. Patients with cancer are a vulnerable population who may be facing imminent death, body changes, and other psychosocial harms[401]. Thus, external influences that affect medical decision making become less ethical and more harmful. Efforts to mitigate FCOI require reforms at multiple levels. Journals and CPG organizations must enforce FCOI policy and adhere to best practices[88]. Institutions must reward unbiased scientific practice[402]. Until the patient is the only external influence that affects clinical decision making, oncology research and practice may fail to capture patient values and preferences for cancer treatment.

Overall, there have been a number of significant advances in the treatment of cancer in the United States and around the globe. Identification of novel mechanisms or signaling pathways and drugs that target those mechanisms or pathways have resulted in dozens of new therapies for patients in previous decades. Knowing the true effect that these interventions have on patient outcomes is contingent on the rigor and reproducibility of oncology evidence. Whether the oncology community continues to reinforce a commitment to robust critical appraisal of new evidence is contingent on its ability to remain unbiased from external influences, such as financial relationships with drug firms. Improvements in research reporting, data availability, selection of trial endpoint, solicitation of patient values, and drug approval regulations will all be critical to the future success of oncology research. The success of cancer treatment should not be measured by the number of treatment options or degree of hype around novel therapies, but rather by whether patients are informed, understood, and empowered to trust that oncologists work for them in all facets of their research and clinical career.

# REFERENCES

1. American Cancer Society. Lifetime Risk of Developing or Dying From Cancer. Published January 13, 2020. Accessed February 15, 2021. https://www.cancer.org/cancer/cancer-basics/lifetime-probability-of-developing-or-dying-from-cancer.html

2. National Cancer Institute. Surveillance, Epidemiology, and End Results Program. Published 2021. Accessed February 15, 2021. https://seer.cancer.gov/explorer/application.html?site=1&data_type=2&graph_type=2&compareBy=sex&chk_sex_3=3&chk_sex_2=2&race=1&age_range=1&advopt_precision=1&advopt_display=2

3. Beaver JA, Howie LJ, Pelosof L, et al. A 25-Year Experience of US Food and Drug Administration Accelerated Approval of Malignant Hematology and Oncology Drugs and Biologics: A Review. *JAMA Oncol*. Published online March 1, 2018. doi:10.1001/jamaoncol.2017.5618

4. Welch HG, Mazer BL, Adamson AS. The Rapid Rise in Cutaneous Melanoma Diagnoses. *N Engl J Med*. 2021;384(1):72-79.

5. Davies L, Welch HG. Increasing incidence of thyroid cancer in the United States, 1973-2002. *JAMA*. 2006;295(18):2164-2167.

6. Collado L, Brownell I. The crippling financial toxicity of cancer in the United States. *Cancer Biol Ther*. 2019;20(10):1301-1303.

7. Huntington SF, Davidoff AJ, Gross CP. Precision Medicine in Oncology II: Economics of Targeted Agents and Immuno-Oncology Drugs. *J Clin Oncol*. 2020;38(4):351-358.

8. Ramsey SD, Dusetzina SB. Weighing Costs and Benefits in the Economics of Cancer Care. *J Clin Oncol*. 2020;38(4):289-291.

9. Green AK, Ohn JA, Bach PB. Review of current policy strategies to reduce US cancer drug costs. *J Clin Oncol*. 2020;38(4):372.

10. Ioannidis JPA. Why most published research findings are false. *PLoS Med*. 2005;2(8):e124.

11. Murad MH, Asi N, Alsawas M, Alahdab F. New evidence pyramid. *Evid Based Med*. 2016;21(4):125-127.

12. Ransohoff DF. Bias as a threat to the validity of cancer molecular-marker research. *Nat Rev Cancer*. 2005;5(2):142-149.

13. Chubak J, Boudreau DM, Wirtz HS, McKnight B, Weiss NS. Threats to validity of nonrandomized studies of postdiagnosis exposures on cancer recurrence and survival. *J Natl Cancer Inst*. 2013;105(19):1456-1462.

14. Vucic EA, Thu KL, Robison K, et al. Translating cancer "omics" to improved outcomes. *Genome Res*. 2012;22(2):188-195.

15. Mbuagbaw L, Lawson DO, Puljak L, Allison DB, Thabane L. A tutorial on methodological studies: the what, when, how and why. *BMC Med Res Methodol*. 2020;20(1):226.

16. National Cancer Institute. NCI Dictionary of Cancer Terms. Published 2021. Accessed February 15, 2021. https://www.cancer.gov/publications/dictionaries/cancer-terms/expand/C

17. Higgins JPT, Thomas J, Chandler J, et al. *Cochrane Handbook for Systematic Reviews of Interventions, 2nd Edition*. Chichester (UK): John Wiley & Sons; 2019.

18. Freedman LP, Cockburn IM, Simcoe TS. The Economics of Reproducibility in Preclinical Research. *PLoS Biol*. 2015;13(6):e1002165.

19. Cabral BP, Fonseca M da GD, Mota FB. The recent landscape of cancer research worldwide: a bibliometric and network analysis. *Oncotarget*. 2018;9(55):30474.

20. Cancer Action Network. NIH and NCI Funding Increased in FY 2021 Budget Deal; Access to Care Prioritized. Published December 21, 2020. Accessed February 15, 2021. https://www.fightcancer.org/releases/nih-and-nci-funding-increased-fy-2021-budget-deal-access-care-prioritized

21. Pereira TV, Horwitz RI, Ioannidis JPA. Empirical evaluation of very large treatment effects of medical interventions. *JAMA*. 2012;308(16):1676-1684.

22. Cunningham A. *A Highest Fall Survived without a Parachute*. London (UK): Guiness Book of World Records; 2002.

23. British Parachute Association. How Safe? Published January 31, 2020. Accessed February 8, 2021. https://britishskydiving.org/how-safe/

24. Rosenberger WF, Lachin JM. *Randomization in Clinical Trials: Theory and Practice*. John Wiley & Sons; 2015.

25. Day SJ, Altman DG. Blinding in clinical trials and other studies. *BMJ*. 2000;321(7259):504.

26. Noseworthy JH, Ebers GC, Vandervoort MK, Farquhar RE, Yetisir E, Roberts R. The impact of blinding on the results of a randomized, placebo-controlled multiple sclerosis clinical trial. *Neurology*. 1994;44(1):16-20.

27. Herson J. *Data and Safety Monitoring Committees in Clinical Trials*. CRC Press; 2016.

28. Institute of Medicine, Board on Health Sciences Policy, Forum on Drug Discovery, Development, and Translation. *Transforming Clinical Research in the United States: Challenges and Opportunities: Workshop Summary*. National Academies Press; 2010.

29. DeAngelis CD, Drazen JM, Frizelle FA, et al. Clinical trial registration: a statement from the International Committee of Medical Journal Editors. *JAMA*. 2004;292(11):1363-1364.

30. Ballantyne A. How should we think about clinical data ownership? *J Med Ethics*. 2020;46(5):289-294.

31. Hopkins AM, Rowland A, Sorich MJ. Data sharing from pharmaceutical industry sponsored clinical studies: audit of data availability. *BMC Med*. 2018;16(1):165.

32. Food and Drug Administration. Title VIII—Clinical Trial Database Sec. 801. Expanded Clinical Trial Registry Data Bank. Published September 27, 2007. Accessed April 10, 2019. https://www.govinfo.gov/content/pkg/PLAW-110publ85/pdf/PLAW-110publ85.pdf#page=82

33. Zarin DA, Tse T, Williams RJ, Rajakannan T. Update on Trial Registration 11 Years after the ICMJE Policy Was Established. *N Engl J Med*. 2017;376(4):383-391.

34. Home - ClinicalTrials.gov. Accessed November 24, 2020. https://clinicaltrials.gov/

35. Huić M, Marušić M, Marušić A. Completeness and changes in registered data and reporting bias of randomized controlled trials in ICMJE journals after trial registration policy. *PLoS One*. 2011;6(9):e25258.

36. Wayant C, Scheckel C, Hicks C, et al. Evidence of selective reporting bias in hematology journals: A systematic review. *PLoS One*. 2017;12(6):e0178379.

37. Ioannidis JPA, Fanelli D, Dunne DD, Goodman SN. Meta-research: Evaluation and Improvement of Research Methods and Practices. *PLoS Biol*. 2015;13(10):e1002264.

38. Palpacuer C, Hammas K, Duprez R, Laviolle B, Ioannidis JPA, Naudet F. Vibration of effects from diverse inclusion/exclusion criteria and analytical choices: 9216 different ways to perform an indirect comparison meta-analysis. *BMC Med*. 2019;17(1):174.

39. Hilal T, Sonbol MB, Prasad V. Analysis of Control Arm Quality in Randomized Clinical Trials Leading to Anticancer Drug Approval by the US Food and Drug Administration. *JAMA Oncol*. 2019;5(6):887-892.

40. Wayant C, Walters C, Zaaza Z, et al. Evaluation of financial conflicts of interest among physician-authors of American College of Rheumatology clinical practice guidelines. *Arthritis & Rheumatology*. Published online 2020. doi:10.1002/art.41224

41. Mitchell AP, Basch EM, Dusetzina SB. Financial Relationships With Industry Among National Comprehensive Cancer Network Guideline Authors. *JAMA Oncol*. 2016;2(12):1628-1631.

42. Wayant C, Puljak L, Bibens M, Vassar M. Risk of Bias and Quality of Reporting in Colon and Rectal Cancer Systematic Reviews Cited by National Comprehensive Cancer Network Guidelines. *J Gen Intern Med*. Published online January 16, 2020. doi:10.1007/s11606-020-05639-y

43. Nissen T, Wayant C, Wahlstrom A, et al. Methodological quality, completeness of reporting and use of systematic reviews as evidence in clinical practice guidelines for paediatric overweight and obesity. *Clin Obes*. 2017;7(1):34-45.

44. NCCN Disclosure Policies and Potential Conflicts of Interest. Accessed November 10, 2017. https://www.nccn.org/about/disclosure.aspx

45. Abola MV, Prasad V. The Use of Superlatives in Cancer Research. *JAMA Oncol*. 2016;2(1):139-141.

46. National Cancer Institute. Cancer Moonshot. Accessed March 17, 2018. https://www.cancer.gov/research/key-initiatives/moonshot-cancer-initiative

47. Hampson LA, Joffe S, Fowler R, Verter J, Emanuel EJ. Frequency, type, and monetary value of financial conflicts of interest in cancer clinical research. *J Clin Oncol*. 2007;25(24):3609-3614.

48. Wagner J, Marquart J, Ruby J, et al. Frequency and level of evidence used in recommendations by the National Comprehensive Cancer Network guidelines beyond approvals of the US Food and Drug Administration: retrospective observational study. *BMJ*. 2018;360. https://www.bmj.com/content/360/bmj.k668.full

49. Prasad V. Double-crossed: why crossover in clinical trials may be distorting medical science. *J Natl Compr Canc Netw*. 2013;11(5):625-627.

50. Wayant C, Ross A, Vassar M. Methodological quality of oncology noninferiority clinical trials. *Crit Rev Oncol Hematol*. 2020;149:102938.

51. Wayant C, Margalski D, Vaughn K, Vassar M. Evaluation of spin in oncology clinical trials. *Crit Rev Oncol Hematol*. 2019;144:102821.

52. Boutron I, Altman DG, Hopewell S, Vera-Badillo F, Tannock I, Ravaud P. Impact of spin in the abstracts of articles reporting results of randomized controlled trials in the field of cancer: the SPIIN randomized controlled trial. *J Clin Oncol*. 2014;32(36):4120-4126.

53. Booth CM, Eisenhauer EA. Progression-free survival: meaningful or simply measurable? *J Clin Oncol*. 2012;30(10):1030-1033.

54. Williamson PR, Altman DG, Bagley H, et al. The COMET Handbook: version 1.0. *Trials*. 2017;18(Suppl 3):280.

55. Frank L, Basch E, Selby JV, Patient-Centered Outcomes Research Institute. The PCORI perspective on patient-centered outcomes research. *JAMA*. 2014;312(15):1513-1514.

56. Canadian Institutes of Health Research. Canada's Strategy for Patient-Oriented Research. Published August 2011. Accessed December 7, 2020. https://cihr-irsc.gc.ca/e/44000.html

57. Chen EY, Joshi SK, Tran A, Prasad V. Estimation of Study Time Reduction Using Surrogate End Points Rather Than Overall Survival in Oncology Clinical Trials. *JAMA Intern Med*. 2019;179(5):642-647.

58. Raphael MJ, Robinson A, Booth CM, et al. The Value of Progression-Free Survival as a Treatment End Point Among Patients With Advanced Cancer: A Systematic Review and Qualitative Assessment of the Literature. *JAMA Oncol*. Published online September 26, 2019. doi:10.1001/jamaoncol.2019.3338

59. Schwartz LH, Bogaerts J, Ford R, et al. Evaluation of lymph nodes with RECIST 1.1. *Eur J Cancer*. 2009;45(2):261-267.

60. Hansen RP, Vedsted P, Sokolowski I, Søndergaard J, Olesen F. Time intervals from first symptom to treatment of cancer: a cohort study of 2,212 newly diagnosed cancer patients. *BMC Health Serv Res*. 2011;11:284.

61. Prasad V, Kim C, Burotto M, Vandross A. The Strength of Association Between Surrogate End Points and Survival in Oncology: A Systematic Review of Trial-Level Meta-analyses. *JAMA Intern Med*. 2015;175(8):1389-1398.

62. Haslam A, Hey SP, Gill J, Prasad V. A systematic review of trial-level meta-analyses measuring the strength of association between surrogate end-points and overall survival in oncology. *Eur J Cancer*. 2019;106:196-211.

63. Kovic B, Jin X, Kennedy SA, et al. Evaluating Progression-Free Survival as a Surrogate Outcome for Health-Related Quality of Life in Oncology: A Systematic Review and Quantitative Analysis. *JAMA Intern Med*. 2018;178(12):1586-1596.

64. Hwang TJ, Gyawali B. Association between progression-free survival and patients' quality of life in cancer clinical trials. *International journal of cancer*. 2019;144(7):1746-1751.

65. Taichman DB, Sahni P, Pinborg A, et al. Data sharing statements for clinical trials. *BMJ*. 2017;357:j2372.

66. Krumholz HM, Ross JS. A model for dissemination and independent analysis of industry data. *JAMA*. 2011;306(14):1593-1594.

67. Mello MM, Francer JK, Wilenzick M, Teden P, Bierer BE, Barnes M. Preparing for Responsible Sharing of Clinical Trial Data. *New England Journal of Medicine*. 2013;369(17):1651-1658. doi:10.1056/nejmhle1309073

68. Whitty CJM, Mundel T, Farrar J, Heymann DL, Davies SC, Walport MJ. Providing incentives to share data early in health emergencies: the role of journal editors. *Lancet*. 2015;386(10006):1797.

69. Tenopir C, Dalton ED, Allard S, et al. Changes in Data Sharing and Data Reuse Practices and Perceptions among Scientists Worldwide. *PLoS One*. 2015;10(8):e0134826.

70. Gorgolewski KJ, Margulies DS, Milham MP. Making data sharing count: a publication-based solution. *Front Neurosci*. 2013;7:9.

71. Roberts D, Dalziel S. Antenatal corticosteroids for accelerating fetal lung maturation for women at risk of preterm birth. *Cochrane Database Syst Rev*. 2006;(3):CD004454.

72. International prospective register of systematic reviews. PROSPERO. Accessed December 7, 2020. https://www.crd.york.ac.uk/prospero/

73. Rombey T, Doni K, Hoffmann F, Pieper D, Allers K. More systematic reviews were registered in PROSPERO each year, but few records' status was up-to-date. *J Clin Epidemiol*. 2020;117:60-67.

74. Gopalakrishnan S, Ganeshkumar P. Systematic Reviews and Meta-analysis: Understanding the Best Evidence in Primary Healthcare. *J Family Med Prim Care*. 2013;2(1):9-14.

75. Linares-Espinós E, Hernández V, Domínguez-Escrig JL, et al. Methodology of a systematic review. *Actas Urol Esp*. 2018;42(8):499-506.

76. Muka T, Glisic M, Milic J, et al. A 24-step guide on how to design, conduct, and successfully publish a systematic review and meta-analysis in medical research. *Eur J Epidemiol*. 2020;35(1):49-60.

77. Moher D, Liberati A, Tetzlaff J, Altman DG, PRISMA Group. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *PLoS Med*. 2009;6(7):e1000097.

78. Paez A. Gray literature: An important resource in systematic reviews. *J Evid Based Med*. 2017;10(3):233-240.

79. Polanin JR, Pigott TD, Espelage DL, Grotpeter JK. Best practice guidelines for abstract screening large-evidence systematic reviews and meta-analyses. *Res Synth Methods*. 2019;10(3):330-342.

80. Cañón M, Buitrago-Gómez Q. The Research Question in Clinical Practice: A Guideline for Its Formulation. *Rev Colomb Psiquiatr*. 2018;47(3):193-200.

81. Holmes J, Herrmann D, Koller C, et al. Heterogeneity of systematic reviews in oncology. *Proc* . 2017;30(2):163-166.

82. Page MJ, McKenzie JE, Kirkham J, et al. Bias due to selective inclusion and reporting of outcomes and analyses in systematic reviews of randomised trials of healthcare interventions. *Cochrane Database Syst Rev*. 2014;(10):MR000035.

83. Sterne JAC, Savović J, Page MJ, et al. RoB 2: a revised tool for assessing risk of bias in randomised trials. *BMJ*. 2019;366:l4898.

84. Sterne JA, Hernán MA, Reeves BC, et al. ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions. *BMJ*. 2016;355:i4919.

85. Richardson M, Garner P, Donegan S. Interpretation of subgroup analyses in systematic reviews: A tutorial. *Clinical Epidemiology and Global Health*. 2019;7(2):192-198.

86. Lau J, Ioannidis JP, Schmid CH. Quantitative synthesis in systematic reviews. *Ann Intern Med*. 1997;127(9):820-826.

87. Murad MH. Clinical Practice Guidelines: A Primer on Development and Dissemination. *Mayo Clin Proc*. 2017;92(3):423-433.

88. Institute of Medicine Board on Health Care Services Committee on Standards for Developing Trustworthy Clinical Practice Guidelines. *Clinical Practice Guidelines We Can Trust*. National Academies Press; 2011.

89. Centers for Medicare and Medicaid Services. Medicare Coverage Database Compendia. Accessed December 7, 2020. https://www.cms.gov/medicare-coverage-database/indexes/medicare-coverage-documents-index.aspx?MCDIndexType=6&mcdtypename=Compendia

90. Robinson JC. Value-based physician payment in oncology: public and private insurer initiatives. *Milbank Q*. 2017;95(1):184-203.

91. Bandari J, Turner RM 2nd, Jacobs BL, Canes D, Moinzadeh A, Davies BJ. The Relationship of Industry Payments to Prescribing Behavior: A Study of Degarelix and Denosumab. *Urol Pract*. 2017;4(1):14-20.

92. Mitchell AP, Winn AN, Dusetzina SB. Pharmaceutical Industry Payments and Oncologists' Selection of Targeted Cancer Therapies in Medicare Beneficiaries. *JAMA Intern Med*. Published online April 9, 2018. doi:10.1001/jamainternmed.2018.0776

93. Zezza MA, Bachhuber MA. Payments from drug companies to physicians are associated with higher volume and more expensive opioid analgesic prescribing. *PLoS One*. 2018;13(12):e0209383.

94. Chen Y, Yang K, Marušic A, et al. A Reporting Tool for Practice Guidelines in Health Care: The RIGHT Statement. *Ann Intern Med*. 2017;166(2):128-132.

95. Brouwers MC, Kho ME, Browman GP, et al. AGREE II: advancing guideline development, reporting and evaluation in health care. *CMAJ*. 2010;182(18):E839-E842.

96. Guyatt GH, Oxman AD, Vist GE, et al. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. *BMJ*. 2008;336(7650):924-926.

97. Guyatt GH, Oxman AD, Kunz R, et al. GRADE guidelines 6. Rating the quality of evidence—imprecision. *J Clin Epidemiol*. 2011;64(12):1283-1293.

98. Guyatt GH, Oxman AD, Kunz R, et al. GRADE guidelines: 8. Rating the quality of evidence—indirectness. *Journal of Clinical Epidemiology*. 2011;64(12):1303-1310. doi:10.1016/j.jclinepi.2011.04.014

99. Guyatt GH, Oxman AD, Kunz R, et al. GRADE guidelines: 7. Rating the quality of evidence—inconsistency. *Journal of Clinical Epidemiology*. 2011;64(12):1294-1302. doi:10.1016/j.jclinepi.2011.03.017

100. Whiting P, Savović J, Higgins JPT, et al. ROBIS: A new tool to assess risk of bias in systematic reviews was developed. *J Clin Epidemiol*. 2016;69:225-234.

101. Guyatt GH, Oxman AD, Vist G, et al. GRADE guidelines: 4. Rating the quality of evidence—study limitations (risk of bias). *J Clin Epidemiol*. 2011;64(4):407-415.

102. ESMO Clinical Practice Guidelines - Standard Operating Procedures. ESMO Guidelines Methodology. Published October 2017. Accessed November 10, 2017. http://www.esmo.org/Guidelines/ESMO-Guidelines-Methodology

103. Johnson L, Stricker RB. The Infectious Diseases Society of America Lyme guidelines: a cautionary tale about the development of clinical practice guidelines. *Philos Ethics Humanit Med*. 2010;5:9.

104. Kalata P, Martus P, Zettl H, et al. Differences between clinical trial participants and patients in a population-based registry: the German Rectal Cancer Study vs. the Rostock Cancer Registry. *Dis Colon Rectum*. 2009;52(3):425-437.

105. Hutchins LF, Unger JM, Crowley JJ, Coltman CA Jr, Albain KS. Underrepresentation of patients 65 years of age or older in cancer-treatment trials. *N Engl J Med*. 1999;341(27):2061-2067.

106. Hotta K, Ueoka H, Kiura K, Tabata M, Tanimoto M. An overview of 48 elderly-specific clinical trials of systemic chemotherapy for advanced non-small cell lung cancer. *Lung Cancer*. 2004;46(1):61-76.

107. Jennens RR, Giles GG, Fox RM. Increasing underrepresentation of elderly patients with advanced colorectal or non-small-cell lung cancer in chemotherapy trials. *Intern Med J*. 2006;36(4):216-220.

108. Gerber DE, Pruitt SL, Halm EA. Should criteria for inclusion in cancer clinical trials be expanded? *J Comp Eff Res*. 2015;4(4):289-291.

109. Gyawali B, Addeo A. Negative phase 3 randomized controlled trials: Why cancer drugs fail the last barrier? *Int J Cancer*. 2018;143(8):2079-2081.

110. Ledley FD, McCoy SS, Vaughan G, Cleary EG. Profitability of Large Pharmaceutical Companies Compared With Other Large Public Companies. *JAMA*. 2020;323(9):834-843.

111. Unger JM, Cook E, Tai E, Bleyer A. The Role of Clinical Trial Participation in Cancer Research: Barriers, Evidence, and Strategies. *Am Soc Clin Oncol Educ Book*. 2016;35:185-198.

112. Latimer KM. Lung Cancer: Clinical Presentation and Diagnosis. *FP Essent*. 2018;464:23-26.

113. Corner J, Hopkinson J, Fitzsimmons D, Barclay S, Muers M. Is late diagnosis of lung cancer inevitable? Interview study of patients' recollections of symptoms before diagnosis. *Thorax*. 2005;60(4):314-319.

114. Schwartz LH, Litière S, de Vries E, et al. RECIST 1.1-Update and clarification: From the RECIST committee. *Eur J Cancer*. 2016;62:132-137.

115. Gyawali B, D'Andrea E, Franklin JM, Kesselheim AS. A correlation analysis to assess event-free survival as a trial-level surrogate for overall survival in early breast cancer. *EClinicalMedicine*. Published online January 29, 2021:100730.

116. Fojo T, Mailankody S, Lo A. Unintended consequences of expensive cancer therapeutics—the pursuit of marginal indications and a me-too mentality that stifles innovation and creativity: the John Conley Lecture. *JAMA Otolaryngology--Head & Neck Surgery*. 2014;140(12):1225-1236.

117. Uno H, Claggett B, Tian L, et al. Moving beyond the hazard ratio in quantifying the between-group difference in survival analysis. *J Clin Oncol*. 2014;32(22):2380-2385.

118. Uno H, Wittes J, Fu H, et al. Alternatives to Hazard Ratios for Comparing the Efficacy or Safety of Therapies in Noninferiority Studies. *Ann Intern Med*. 2015;163(2):127-134.

119. Pak K, Uno H, Kim DH, et al. Interpretability of Cancer Clinical Trial Results Using Restricted Mean Survival Time as an Alternative to the Hazard Ratio. *JAMA Oncol*. 2017;3(12):1692-1696.

120. Prasad V, Bilal U. The role of censoring on progression free survival: oncologist discretion advised. *Eur J Cancer*. 2015;51(16):2269-2271.

121. Winer A, Bodor JN, Borghaei H. Identifying and managing the adverse effects of immune checkpoint blockade. *J Thorac Dis*. 2018;10(Suppl 3):S480-S489.

122. Rosen K, Prasad V, Chen EY. Censored patients in Kaplan–Meier plots of cancer drugs: An empirical analysis of data sharing. *Eur J Cancer*. 2020;141:152-161.

123. Prasad V, Grady C. The misguided ethics of crossover trials. *Contemp Clin Trials*. 2014;37(2):167-169.

124. Henshall C, Latimer NR, Sansom L, Ward RL. TREATMENT SWITCHING IN CANCER TRIALS: ISSUES AND PROPOSALS. *Int J Technol Assess Health Care*. 2016;32(3):167-174.

125. Haslam A, Prasad V. When is crossover desirable in cancer drug trials and when is it problematic? *Ann Oncol*. 2018;29(5):1079-1081.

126. Todd S, Whitehead A, Stallard N, Whitehead J. Interim analyses and sequential designs in phase III studies. *Br J Clin Pharmacol*. 2001;51(5):394-399.

127. Whitehead J. A unified theory for sequential clinical trials. *Stat Med*. 1999;18(17-18):2271-2286.

128. Woloshin S, Schwartz LM, Bagley PJ, Blunt HB, White B. Characteristics of Interim Publications of Randomized Clinical Trials and Comparison With Final Publications. *JAMA*. 2018;319(4):404-406.

129. Turner EH, Matthews AM, Linardatos E, Tell RA, Rosenthal R. Selective publication of antidepressant trials and its influence on apparent efficacy. *N Engl J Med*. 2008;358(3):252-260.

130. Et T, Heneghan C. Outcome reporting bias. Published 2017. Accessed February 2, 2021. https://catalogofbias.org/biases/outcome-reporting-bias/.

131. FDAAA 801 Requirements. Published September 16, 2016. Accessed August 15, 2017. https://clinicaltrials.gov/ct2/manage-recs/fdaaa

132. Nosek BA, Alter G, Banks GC, et al. SCIENTIFIC STANDARDS. Promoting an open research culture. *Science*. 2015;348(6242):1422-1425.

133. Wiebe J, Detten G, Scheckel C, et al. The heart of the matter: Outcome reporting bias and registration status in cardio-thoracic surgery. *Int J Cardiol*. 2017;227:299-304.

134. Ross A, George D, Wayant C, Hamilton T, Vassar M. Registration Practices of Randomized Clinical Trials in Rhinosinusitis: A Cross-sectional Review. *JAMA Otolaryngol Head Neck Surg*. Published online March 28, 2019. doi:10.1001/jamaoto.2019.0145

135. Vassar M, Roberts W, Cooper CM, Wayant C, Bibens M. Evaluation of selective outcome reporting and trial registration practices among addiction clinical trials. *Addiction*. 2020;115(6):1172-1179.

136. Raghav KPS, Mahajan S, Yao JC, et al. From Protocols to Publications: A Study in Selective Reporting of Outcomes in Randomized Trials in Oncology. *J Clin Oncol*. 2015;33(31):3583-3590.

137. Boutron I, Dutton S, Ravaud P, Altman DG. Reporting and interpretation of randomized controlled trials with statistically nonsignificant results for primary outcomes. *JAMA*. 2010;303(20):2058-2064.

138. Yavchitz A, Boutron I, Bafeta A, et al. Misrepresentation of randomized controlled trials in press releases and news coverage: a cohort study. *PLoS Med*. 2012;9(9):e1001308.

139. Lazarus C, Haneef R, Ravaud P, Hopewell S, Altman DG, Boutron I. Peer reviewers identified spin in manuscripts of nonrandomized studies assessing therapeutic interventions, but their impact on spin in abstract conclusions was limited. *J Clin Epidemiol*. 2016;77:44-51.

140. Grzybowski A, Kanclerz P. Language Bias and Methodological Issues in Determining Reliable Evidence for Systematic Reviews. *JAMA Ophthalmol*. 2019;137(1):118-119.

141. Sterne J, Egger M, Moher D. 10.2.2.3 Citation bias. In: Higgins JPT GS, ed. *Cochrane Handbook for Systematic Reviews of Interventions Version 5.1.0*. The Cochrane Collaboration; 2011.

142. Moher D, Tetzlaff J, Tricco AC, Sampson M, Altman DG. Epidemiology and Reporting Characteristics of Systematic Reviews. *PLoS Medicine*. 2007;4(3):e78. doi:10.1371/journal.pmed.0040078.

143. Page MJ, Altman DG, Shamseer L, et al. Reproducible research practices are underused in systematic reviews of biomedical interventions. *J Clin Epidemiol*. 2018;94:8-18.

144. Egger M, Davey Smith G, Schneider M, Minder C. Bias in meta-analysis detected by a simple, graphical test. *BMJ*. 1997;315(7109):629-634.

145. Runjic E, Rombey T, Pieper D, Puljak L. Half of systematic reviews about pain registered in PROSPERO were not published and the majority had inaccurate status. *J Clin Epidemiol*. 2019;116:114-121.

146. Flores H, Kannan D, Ottwell R, et al. Evaluation of spin in the abstracts of systematic reviews and meta-analyses on breast cancer treatment, screening, and quality of life outcomes: A cross-sectional study. *Journal of Cancer Policy*. 2021;27:100268. doi:10.1016/j.jcpo.2020.100268

147. Jones C, Rulon Z, Arthur W, et al. Evaluation of spin in the abstracts of systematic reviews and meta-analyses related to the treatment of proximal humerus fractures. *J Shoulder Elbow Surg*. Published online January 19, 2021. doi:10.1016/j.jse.2020.11.026

148. Okonya O, Lai E, Khattab M, et al. Evaluation of Spin in the Abstracts of Systematic Reviews and Meta-analyses of Treatments for Glaucoma. *J Glaucoma*. 2020;Publish Ahead of Print. doi:10.1097/IJG.0000000000001735

149. Romeo V, Stanzione A, Gaudieri V, et al. A critical appraisal of the quality of 18F-FDG PET/CT guidelines in oncology using the AGREE II tool: A EuroAIM initiative. *Eur J Radiol*. 2020;126:108930.

150. Irajpour A, Hashemi M, Taleghani F. The quality of guidelines on the end-of-life care: a systematic quality appraisal using AGREE II instrument. *Support Care Cancer*. 2020;28(4):1555-1561.

151. Wayant C, Cooper C, Turner D ’arcy, Vassar M. Evaluation of the NCCN guidelines using the RIGHT Statement and AGREE-II instrument: a cross-sectional review. *BMJ Evid Based Med*. 2019;24(6):219-226.

152. Rosenbaum L. Beyond Moral Outrage — Weighing the Trade-Offs of COI Regulation. *N Engl J Med*. 2015;372(21):2064-2068.

153. Tibau A, Bedard PL, Srikanthan A, et al. Author financial conflicts of interest, industry funding, and clinical practice guidelines for anticancer drugs. *J Clin Oncol*. 2015;33(1):100-106.

154. Carey C, Lieber EMJ, Miller S. Drug Firms’ Payments and Physicians’ Prescribing Behavior in Medicare Part D. *Natl Bur Econ Res Bull Aging Health*. Published online February 2020. http://www-personal.umich.edu/~mille/CareyLieberMiller_PhysicianPayments2015.pdf

155. Lerner TG, Miranda M da C, Lera AT, et al. The prevalence and influence of self-reported conflicts of interest by editorial authors of phase III cancer trials. *Contemp Clin Trials*. 2012;33(5):1019-1022.

156. Bonnot B, Yavchitz A, Mantz J, Paugam-Burtz C, Boutron I. Selective primary outcome reporting in high-impact journals of anaesthesia and pain. *Br J Anaesth*. 2016;117(4):542-543.

157. Simera I, Moher D, Hirst A, Hoey J, Schulz KF, Altman DG. Transparent and accurate reporting increases reliability, utility, and impact of your research: reporting guidelines and the EQUATOR Network. *BMC Med*. 2010;8:24.

158. Glasziou P, Meats E, Heneghan C, Shepperd S. What is missing from descriptions of treatment in trials and reviews? *BMJ*. 2008;336(7659):1472-1474.

159. Helfer B, Prosser A, Samara MT, et al. Recent meta-analyses neglect previous systematic reviews and meta-analyses about the same topic: a systematic examination. *BMC Med*. 2015;13:82.

160. Simera I, Moher D, Hoey J, Schulz KF, Altman DG. TheEQUATOR Network and reporting guidelines: helping toachieve high standards in reporting health research studies. *Maturitas*. 2009;63:4-6.

161. Grellety T, Petit-Monéger A, Diallo A, Mathoulin-Pelissier S, Italiano A. Quality of reporting of phase II trials: a focus on highly ranked oncology journals. *Ann Oncol*. 2014;25(2):536-541.

162. Maillet D, Blay JY, You B, Rachdi A, Gan HK, Péron J. The reporting of adverse events in oncology phase III trials: a comparison of the current status versus the expectations of the EORTC members. *Ann Oncol*. 2016;27(1):192-198.

163. Schulz KF, Altman DG, Moher D, CONSORT Group. CONSORT 2010 Statement: updated guidelines for reporting parallel group randomised trials. *Trials*. 2010;11:32.

164. Péron J, Pond GR, Gan HK, et al. Quality of reporting of modern randomized controlled trials in medical oncology: a systematic review. *J Natl Cancer Inst*. 2012;104(13):982-989.

165. Panic N, Leoncini E, de Belvis G, Ricciardi W, Boccia S. Evaluation of the endorsement of the preferred reporting items for systematic reviews and meta-analysis (PRISMA) statement on the quality of published systematic review and meta-analyses. *PLoS One*. 2013;8(12):e83138.

166. Viergever RF, Ghersi D. The quality of registration of clinical trials. *PLoS One*. 2011;6(2):e14701.

167. Viergever RF, Karam G, Reis A, Ghersi D. The quality of registration of clinical trials: still a problem. *PLoS One*. 2014;9(1):e84727.

168. Su C-X, Han M, Ren J, et al. Empirical evidence for outcome reporting bias in randomized clinical trials of acupuncture: comparison of registered records and subsequent publications. *Trials*. 2015;16:28.

169. Clinical Trial Registration. International Committee of Medical Journal Editors (ICMJE). Accessed October 20, 2017. http: //www.icmje.org/recommendations/browse/publishing-and-editorial-issues/ clinical-trial-registration.html.

170. *Regulation (EU) No 536/2014 of the European Parliament and of the Council of 16 April 2014 on Clinical Trials on Medicinal Products for Human Use, and Repealing Directive 2001/20/EC*.

171. FDAAA for NIH Grantees: The Basics. Accessed May 3, 2017. https://grants.nih.gov/clinicaltrials_fdaaa/the-basics.htm#whatisFDAAA

172. Lang TA, Altman DG. Basic statistical reporting for articles published in biomedical journals: the "Statistical Analyses and Methods in the Published Literature" or the SAMPL Guidelines. *Int J Nurs Stud*. 2015;52(1):5-9.

173. Wayant C, Smith C, Sims M, Vassar M. Hematology journals do not sufficiently adhere to reporting guidelines: a systematic review. *J Thromb Haemost*. 2017;15(4):608-617.

174. Sims MT, Henning NM, Wayant CC, Vassar M. Do emergency medicine journals promote trial registration and adherence to reporting guidelines? A survey of "Instructions for Authors." *Scand J Trauma Resusc Emerg Med*. 2016;24(1):137.

175. Dillman DA, Smyth JD, Christian LM. *Internet, Phone, Mail, and Mixed-Mode Surveys: The Tailored Design Method*. John Wiley & Sons; 2014.

176. McKibbon KA, Wilczynski NL, Haynes RB, Hedges Team. Retrieving randomized controlled trials from medline: a comparison of 38 published search filters. *Health Info Libr J*. 2009;26(3):187-202.

177. Equator Network. Accessed April 17, 2017. https://www.equator-network.org/

178. Sims MT, Checketts JX, Wayant C, Vassar M. Requirements for trial registration and adherence to reporting guidelines in critical care journals: a meta-epidemiological study of journals' instructions for authors. *Int J Evid Based Healthc*. Published online August 31, 2017. doi:10.1097/XEB.0000000000000120

179. Zer A, Prince RM, Amir E, Razak ARA. Forty years of randomized trials in advanced/metastatic soft tissue sarcoma (STS): Endpoint selection, surrogacy and quality of reporting. *J Clin Oncol*. 2015;33(15_suppl):10513-10513.

180. Chelsea Koller, Jonathan Holmes, David Herrmann, Matt Vassar. Quality of systematic review and meta-analysis abstracts in oncology journals. *Cancer Treatment and Research Communications*. 2016;9:70-74.

181. Shanthi Sivendran, Kristina Braine Newport, Adam Albert, Matt D. Galsky. Reporting quality of abstracts in cancer clinical trials. *J Clin Oncol*. 2014;32(15_suppl):6584.

182. Chhapola V, Tiwari S, Brar R, Kanwal SK. An interrupted time series analysis showed suboptimal improvement in reporting quality of trial abstract. *J Clin Epidemiol*. 2016;71:11-17.

183. Berwanger O, Ribeiro RA, Finkelsztejn A, et al. The quality of reporting of trial abstracts is suboptimal: survey of major general medical journals. *J Clin Epidemiol*. 2009;62(4):387-392.

184. Bariani GM, de Celis Ferrari ACR, Precivale M, Arai R, Saad ED, Riechelmann RP. Sample Size Calculation in Oncology Trials: Quality of Reporting and Implications for Clinical Cancer Research. *Am J Clin Oncol*. 2015;38(6):570-574.

185. Vera-Badillo FE, Shapiro R, Ocana A, Amir E, Tannock IF. Bias in reporting of end points of efficacy and toxicity in randomized, clinical trials for women with breast cancer. *Ann Oncol*. 2013;24(5):1238-1244.

186. Sim I, Chan A-W, Gülmezoglu AM, Evans T, Pang T. Clinical trial registration: transparency is the watchword. *Lancet*. 2006;367(9523):1631-1633.

187. The EQUATOR Network (Enhancing the QUAlity and Transparency Of health Research). Accessed August 15, 2017. http://www.equator-network.org/

188. EQUATOR Oncology. Published October 4, 2017. Accessed October 5, 2017. http://www.equator-network.org/library/equator-oncology/

189. You B, Gan HK, Pond G, Chen EX. Consistency in the analysis and reporting of primary end points in oncology randomized controlled trials from registration to publication: a systematic review. *J Clin Oncol*. 2012;30(2):210-216.

190. Hopewell S, Hirst A, Collins GS, Mallett S, Yu L-M, Altman DG. Reporting of participant flow diagrams in published reports of randomized trials. *Trials*. 2011;12:253.

191. Toulmonde M, Bellera C, Mathoulin-Pelissier S, Debled M, Bui B, Italiano A. Quality of randomized controlled trials reporting in the treatment of sarcomas. *J Clin Oncol*. 2011;29(9):1204-1209.

192. Bhide A, Shah PS, Acharya G. A simplified guide to randomized controlled trials. *Acta Obstet Gynecol Scand*. 2018;97(4):380-387.

193. Wood L, Egger M, Gluud LL, et al. Empirical evidence of bias in treatment effect estimates in controlled trials with different interventions and outcomes: meta-epidemiological study. *BMJ*. 2008;336(7644):601-605.

194. Schulz KF, Chalmers I, Hayes RJ, Altman DG. Empirical evidence of bias. Dimensions of methodological quality associated with estimates of treatment effects in controlled trials. *JAMA*. 1995;273(5):408-412.

195. Huang YQ, Traore K, Ibrahim B, Sewitch MJ, Nguyen LHP. Reporting quality of randomized controlled trials in otolaryngology: review of adherence to the CONSORT statement. *J Otolaryngol Head Neck Surg*. 2018;47(1):34.

196. Münter NH, Stevanovic A, Rossaint R, Stoppe C, Sanders RD, Coburn M. CONSORT item adherence in top ranked anaesthesiology journals in 2011: a retrospective analysis. *Eur J Anaesthesiol*. 2015;32(2):117-125.

197. Altman DG, Doré CJ. Randomisation and baseline comparisons in clinical trials. *Lancet*. 1990;335(8682):149-153.

198. Williams DH, Davis CE. Reporting of assignment methods in clinical trials. *Control Clin Trials*. 1994;15(4):294-298.

199. Schulz KF, Chalmers I, Altman DG, Grimes DA, Doré CJ. The methodologic quality of randomization as assessed from reports of trials in specialist and general medical journals. *Online J Curr Clin Trials*. 1995;Doc No 197:[81 paragraphs].

200. Glasziou P, Altman DG, Bossuyt P, et al. Reducing waste from incomplete or unusable reports of biomedical research. *Lancet*. 2014;383(9913):267-276.

201. Duff JM, Leather H, Walden EO, LaPlant KD, George TJ Jr. Adequacy of published oncology randomized controlled trials to provide therapeutic details needed for clinical application. *J Natl Cancer Inst*. 2010;102(10):702-705.

202. Goldstein DA, Stemmer SM, Gordon N. The cost and value of cancer drugs – are new innovations outpacing our ability to pay? *Israel Journal of Health Policy Research*. 2016;5(1). doi:10.1186/s13584-016-0097-0

203. Kim C, Prasad V. Cancer Drugs Approved on the Basis of a Surrogate End Point and Subsequent Overall Survival: An Analysis of 5 Years of US Food and Drug Administration Approvals. *JAMA Intern Med*. 2015;175(12):1992-1994.

204. Wayant C, Vassar M. A comparison of matched interim analysis publications and final analysis publications oncology clinical trials. *Ann Oncol*. Published online October 11, 2018. doi:10.1093/annonc/mdy447

205. Food and Drug Administration (FDA). Hematology/Oncology (Cancer) Approvals & Safety Notifications. Published February 28, 2019. Accessed March 7, 2019. https://www.fda.gov/drugs/informationondrugs/approveddrugs/ucm279174.htm

206. Hoffmann TC, Glasziou PP, Boutron I, et al. Better reporting of interventions: template for intervention description and replication (TIDieR) checklist and guide. *BMJ*. 2014;348:g1687.

207. Wayant CC. Reporting of clinical trial interventions for recent FDA-approved anticancer medications. Published online June 18, 2019. Accessed September 10, 2019. https://osf.io/6ujzn/

208. Yamato TP, Maher CG, Saragiotto BT, Catley MJ, Moseley AM. Rasch analysis suggested that items from the template for intervention description and replication (TIDieR) checklist can be summed to create a score. *J Clin Epidemiol*. 2018;101:28-34.

209. Flinn IW, Hillmen P, Montillo M, et al. The phase 3 DUO trial: duvelisib vs ofatumumab in relapsed and refractory CLL/SLL. *Blood*. 2018;132(23):2446-2455.

210. Drilon A, Laetsch TW, Kummar S, et al. Efficacy of Larotrectinib in TRK Fusion–Positive Cancers in Adults and Children. *N Engl J Med*. 2018;378(8):731-739.

211. Mok TS, Wu Y-L, Ahn M-J, et al. Osimertinib or Platinum-Pemetrexed in EGFR T790M-Positive Lung Cancer. *N Engl J Med*. 2017;376(7):629-640.

212. Birkmeyer NJO, Goodney PP, Stukel TA, Hillner BE, Birkmeyer JD. Do cancer centers designated by the National Cancer Institute have better surgical outcomes? *Cancer*. 2005;103(3):435-441.

213. Cheung MC, Hamilton K, Sherman R, et al. Impact of teaching facility status and high-volume centers on outcomes for lung cancer resection: an examination of 13,469 surgical patients. *Ann Surg Oncol*. 2009;16(1):3-13.

214. Lou Y, Dholaria B, Soyano A, et al. Survival trends among non-small-cell lung cancer patients over a decade: impact of initial therapy at academic centers. *Cancer Med*. 2018;7(10):4932-4942.

215. Pfister DG, Rubin DM, Elkin EB, et al. Risk Adjusting Survival Outcomes in Hospitals That Treat Patients With Cancer Without Information on Cancer Stage. *JAMA Oncol*. 2015;1(9):1303-1310.

216. Dimond EP, St. Germain D, Nacpil LM, et al. Creating a "culture of research" in a community hospital: Strategies and tools from the National Cancer Institute Community Cancer Centers Program. *Clin Trials*. 2015;12(3):246-256.

217. Warnecke RB, Johnson TP, Kaluzny AD, Ford LG. The community clinical oncology program: its effect on clinical practice. *Jt Comm J Qual Improv*. 1995;21(7):336-339.

218. Minasian LM, Carpenter WR, Weiner BJ, et al. Translating research into evidence-based practice: the National Cancer Institute Community Clinical Oncology Program. *Cancer*. 2010;116(19):4440-4449.

219. Chen TW, Razak AR, Bedard PL, Siu LL, Hansen AR. A systematic review of immune-related adverse event reporting in clinical trials of immune checkpoint inhibitors. *Ann Oncol*. 2015;26(9):1824-1829.

220. Gedye C, van der Westhuizen A, John T. Checkpoint immunotherapy for cancer: superior survival, unaccustomed toxicities. *Intern Med J*. 2015;45(7):696-701.

221. Di Giacomo AM, Biagioli M, Maio M. The emerging toxicity profiles of anti-CTLA-4 antibodies across clinical indications. *Semin Oncol*. 2010;37(5):499-507.

222. Institute of Medicine (US) Committee on Conflict of Interest in Medical Research, Education, and Practice. *Conflict of Interest in Medical Research, Education, and Practice*. (Lo B, Field MJ, eds.). National Academies Press (US); 2010.

223. DeJong C, Aguilar T, Tseng C-W, Lin GA, Boscardin WJ, Dudley RA. Pharmaceutical Industry-Sponsored Meals and Physician Prescribing Patterns for Medicare Beneficiaries. *JAMA Intern Med*. 2016;176(8):1114-1122.

224. Lundh A, Lexchin J, Mintzes B, Schroll JB, Bero L. Industry sponsorship and research outcome. *Cochrane Database Syst Rev*. 2017;2:MR000033.

225. Office. USGP. The Patient Protection and Affordable Care Act. Public law 111–148. Published March 30, 2010. Accessed December 12, 2017. https://www.gpo.gov/fdsys/pkg/PLAW-111publ148/pdf/PLAW-111publ148.pdf

226. Office USGP. Health Care and Education Reconciliation Act of 2010. Public Law 111–152. Published March 30, 2010. Accessed December 12, 2017. https://www.gpo.gov/fdsys/pkg/PLAW-111publ152/pdf/PLAW-111publ152.pdf

227. Tringale KR, Marshall D, Mackey TK, Connor M, Murphy JD, Hattangadi-Gluth JA. Types and Distribution of Payments From Industry to Physicians in 2015. *JAMA*. 2017;317(17):1774-1784.

228. Hui D, Reddy A, Parsons HA, Bruera E. Reporting of funding sources and conflict of interest in the supportive and palliative oncology literature. *J Pain Symptom Manage*. 2012;44(3):421-430.

229. Kesselheim AS, Lee JL, Avorn J, Servi A, Shrank WH, Choudhry NK. Conflict of interest in oncology publications: a survey of disclosure policies and statements. *Cancer*. 2012;118(1):188-195.

230. US Department of Health & Human Services. Basic HHS Policy for Protection of Human Research Subjects. Title 45 Code of Federal Regulations part 46. Published January 15, 2009. Accessed December 12, 2017. https://www.hhs.gov/ohrp/regulations-and-policy/regulations/45-cfr-46/.

231. Jagsi R, Sheets N, Jankovic A, Motomura AR, Amarnath S, Ubel PA. Frequency, nature, effects, and correlates of conflicts of interest in published clinical cancer research. *Cancer*. 2009;115(12):2783-2791.

232. De Jesus-Morales K, Prasad V. Closed Financial Loops: When They Happen in Government, They're Called Corruption; in Medicine, They're Just a Footnote. *Hastings Cent Rep*. 2017;47(3):9-14.

233. Dorsey ER, de Roulet J, Thompson JP, et al. Funding of US biomedical research, 2003-2008. *JAMA*. 2010;303(2):137-143.

234. Rosenbaum L. Understanding bias--the case for careful study. *N Engl J Med*. 2015;372(20):1959-1963.

235. Cappola AR, FitzGerald GA. Confluence, Not Conflict of Interest: Name Change Necessary. *JAMA*. 2015;314(17):1791-1792.

236. Dobson R. Industry sponsored studies twice as likely to have positive conclusions about costs. *BMJ: British Medical Journal*. 2003;327(7422):1006.

237. Barry MJ. Let the Sun Shine In! An Adventure in Open Payments. *Ann Intern Med*. 2018;168(2):151-152.

238. ProPublica. Government Will Withhold One-Third of the Records from Database of Physician Payments. Published August 15, 2014. Accessed March 6, 2018. https://www.propublica.org/article/government-will-withhold-one-third-of-the-records-from-database-of-physicia

239. Centers for Medicare and Medicaid Services. Open Payments System Reopens, Extends Physician Registration and Review Period.

240. Prasad V, Rajkumar SV. Conflict of interest in academic oncology: moving beyond the blame game and forging a path forward. *Blood Cancer J*. 2016;6(11):e489.

241. Checketts JX, Sims MT, Vassar M. Evaluating Industry Payments Among Dermatology Clinical Practice Guidelines Authors. *JAMA Dermatol*. 2017;153(12):1229-1235.

242. Chiu K, Grundy Q, Bero L. "Spin" in published biomedical literature: A methodological systematic review. *PLoS Biol*. 2017;15(9):e2002173.

243. Ravaud A, Motzer RJ, Pandha HS, et al. Adjuvant Sunitinib in High-Risk Renal-Cell Carcinoma after Nephrectomy. *N Engl J Med*. 2016;375(23):2246-2254.

244. Zhang S. Adjuvant Sunitinib in Renal-Cell Carcinoma. *N Engl J Med*. 2017;376(9):893.

245. Gyawali B, Goldstein DA. The US Food and Drug Administration's Approval of Adjuvant Sunitinib for Renal Cell Cancer: A Case of Regulatory Capture? *JAMA Oncol*. 2018;4(5):623-624.

246. Saint S, Christakis DA, Saha S, et al. Journal reading habits of internists. *J Gen Intern Med*. 2000;15(12):881-884.

247. Marcelo A, Gavino A, Isip-Tan IT, et al. A comparison of the accuracy of clinical decisions based on full-text articles and on journal abstracts alone: a study among residents in a tertiary care hospital. *Evid Based Med*. 2013;18(2):48-53.

248. Barry HC, Ebell MH, Shaughnessy AF, Slawson DC, Nietzke F. Family physicians' use of medical abstracts to guide decision making: style or substance? *J Am Board Fam Pract*. 2001;14(6):437-442.

249. Altwairgi AK, Booth CM, Hopman WM, Baetz TD. Discordance between conclusions stated in the abstract and conclusions in the article: analysis of published randomized controlled trials of systemic therapy in lung cancer. *J Clin Oncol*. 2012;30(28):3552-3557.

250. Kemp R, Prasad V. Surrogate endpoints in oncology: when are they acceptable for regulatory and clinical decisions, and are they currently overused? *BMC Med*. 2017;15(1):134.

251. Food And Drug. Clinical Trial Endpoints for the Approval of Non Small Cell Lung Cancer Drugs and Biologics: Guidance for Industry. Published April 2015. Accessed January 23, 2018. https://www.fda.gov/downloads/drugs/guidances/ucm259421.pdf

252. Cheema PK, Burkes RL. Overall survival should be the primary endpoint in clinical trials for advanced non-small-cell lung cancer. *Curr Oncol*. 2013;20(2):e150-e160.

253. Tan A, Porcher R, Crequit P, Ravaud P, Dechartres A. Differences in Treatment Effect Size Between Overall Survival and Progression-Free Survival in Immunotherapy Trials: A Meta-Epidemiologic Study of Trials With Results Posted at ClinicalTrials.gov. *J Clin Oncol*. 2017;35(15):1686-1694.

254. Wayant C. Search strategy. Open Science Framework. Published July 3, 2018. https://osf.io/fzgh8/

255. U.S. Department of Health and Human Services, Food and Drug Administration, Center for Drug Evaluation and Research (CDER), Center for Biologics Evaluation and Research (CBER). Guidance for Industry Clinical Trial Endpoints for the Approval of Cancer Drugs and Biologics. Published May 2007. Accessed February 6, 2018. https://www.fda.gov/downloads/Drugs/.../Guidances/ucm071590.pdf

256. Ouzzani M, Hammady H, Fedorowicz Z, Elmagarmid A. Rayyan—a web and mobile app for systematic reviews. *Syst Rev*. 2016;5(1):210.

257. Djulbegovic B, Kumar A, Magazin A, et al. Optimism bias leads to inconclusive results-an empirical study. *J Clin Epidemiol*. 2011;64(6):583-593.

258. Boutron I, Ravaud P. Misrepresentation and distortion of research in biomedical literature. *Proc Natl Acad Sci U S A*. 2018;115(11):2613-2619.

259. Master Z, Resnik DB. Hype and public trust in science. *Sci Eng Ethics*. 2013;19(2):321-335.

260. Svensson S, Menkes DB, Lexchin J. Surrogate outcomes in clinical trials: a cautionary tale. *JAMA Intern Med*. 2013;173(8):611-612.

261. Fleming TR, DeMets DL. Surrogate end points in clinical trials: are we being misled? *Ann Intern Med*. 1996;125(7):605-613.

262. Yudkin JS, Lipska KJ, Montori VM. The idolatry of the surrogate. *BMJ*. 2011;343:d7995.

263. Shi Q, Sargent DJ. Meta-analysis for the evaluation of surrogate endpoints in cancer clinical trials. *Int J Clin Oncol*. 2009;14(2):102-111.

264. Mailankody S, Prasad V. Five Years of Cancer Drug Approvals: Innovation, Efficacy, and Costs. *JAMA Oncol*. 2015;1(4):539-540.

265. Kay A, Higgins J, Day AG, Meyer RM, Booth CM. Randomized controlled trials in the era of molecular oncology: methodology, biomarkers, and end points. *Ann Oncol*. 2012;23(6):1646-1651.

266. Gutman SI, Piper M, Grant MD, Basch E, Oliansky DM, Aronson N. *Progression-Free Survival: What Does It Mean for Psychological Well-Being or Quality of Life?* Agency for Healthcare Research and Quality (US); 2013.

267. Institute for Quality and Efficiency in Health Care. Validity of surrogate endpoints in oncology: executive summary. Published November 21, 2011. Accessed April 16, 2018. http://www.iqwig.de/download/A10-05_Executive_Summary_v1-1_Surrogate_endpoints_in_oncology.pdf

268. Sugawara S, Oizumi S, Minato K, et al. Randomized phase II study of concurrent versus sequential alternating gefitinib and chemotherapy in previously untreated non-small cell lung cancer with sensitive EGFR mutations: NEJ005/TCOG0902. *Ann Oncol*. 2015;26(5):888-894.

269. Counsell N, Biri D, Fraczek J, Hackshaw A. Publishing interim results of randomised clinical trials in peer-reviewed journals. *Clin Trials*. 2017;14(1):67-77.

270. Wallach JD, Ciani O, Pease AM, et al. Comparison of treatment effect sizes from pivotal and postapproval trials of novel therapeutics approved by the FDA based on surrogate markers of disease: a meta-epidemiological study. *BMC Med*. 2018;16(1):45.

271. Cooperman T. Trends in FDA approval of Specialty Drugs 1990 through 2017. Published December 15, 2017. Accessed March 17, 2018. http://rjhealthsystems.com/2017/12/15/trends-fda-approval-specialty-drugs-1990-q3-2017/

272. Gyawali B, Hey SP, Kesselheim AS. A Comparison of Response Patterns for Progression-Free Survival and Overall Survival Following Treatment for Cancer With PD-1 Inhibitors: A Meta-analysis of Correlation and Differences in Effect Sizes. *JAMA Network Open*. 2018;1(2):e180416-e180416.

273. Oza AM, Cook AD, Pfisterer J, et al. Standard chemotherapy with or without bevacizumab for women with newly diagnosed ovarian cancer (ICON7): overall survival results of a phase 3 randomised trial. *Lancet Oncol*. 2015;16(8):928-936.

274. Nordlinger B, Sorbye H, Glimelius B, et al. Perioperative FOLFOX4 chemotherapy and surgery versus surgery alone for resectable liver metastases from colorectal cancer (EORTC 40983): long-term results of a randomised, controlled, phase 3 trial. *Lancet Oncol*. 2013;14(12):1208-1215.

275. Silvestri G, Pritchard R, Welch HG. Preferences for chemotherapy in patients with advanced non-small cell lung cancer: descriptive study based on scripted interviews. *BMJ*. 1998;317(7161):771-775.

276. Donohue JM, Cevasco M, Rosenthal MB. A Decade of Direct-to-Consumer Advertising of Prescription Drugs. *N Engl J Med*. 2007;357(7):673-681.

277. Mintzes B. Advertising of Prescription-Only Medicines to the Public: Does Evidence of Benefit Counterbalance Harm? Published online March 19, 2012. doi:10.1146/annurev-publhealth-031811-124540

278. Stange KC. Time to ban direct-to-consumer prescription drug marketing. *Ann Fam Med*. 2007;5(2):101-104.

279. Connors AL. Big bad pharma: an ethical analysis of physician-directed and consumer-directed marketing tactics. *Albany Law Rev*. 2009;73(1):243-282.

280. Delbaere M, Smith MC. Health care knowledge and consumer learning: the case of direct-to-consumer drug advertising. *Health Mark Q*. 2006;23(3):9-29.

281. Boden WE, Diamond GA. DTCA for PTCA—crossing the line in consumer health education? *N Engl J Med*. 2008;358(21):2197-2200.

282. Frosch DL, Krueger PM, Hornik RC, Cronholm PF, Barg FK. Creating demand for prescription drugs: a content analysis of television direct-to-consumer advertising. *Ann Fam Med*. 2007;5(1):6-13.

283. Almasi EA, Stafford RS, Kravitz RL, Mansfield PR. What are the public health effects of direct-to-consumer drug advertising? *PLoS Med*. 2006;3(3):e145.

284. Kuehn BM. FDA weighs limits for online ads. *JAMA*. 2010;303(4):311-313.

285. Frosch DL, Grande D, Tarn DM, Kravitz RL. A Decade of Controversy: Balancing Policy With Evidence in the Regulation of Prescription Drug Advertising. *Am J Public Health*. 2010;100(1):24-32.

286. Schnipper LE, Abel GA. Direct-to-Consumer Drug Advertising in Oncology Is Not Beneficial to Patients or Public Health. *JAMA Oncol*. 2016;2(11):1397-1398.

287. Abel GA, Chen K, Taback N, Hassett MJ, Schrag D, Weeks JC. Impact of oncology-related direct-to-consumer advertising: Association with appropriate and inappropriate prescriptions. *Cancer*. 2013;119(5):1065-1072.

288. Kim H. Trouble Spots in Online Direct-to-Consumer Prescription Drug Promotion: A Content Analysis of FDA Warning Letters. *Int J Health Policy Manag*. 2015;4(12):813-821.

289. Kroschinsky F, Stölzel F, von Bonin S, et al. New drugs, new toxicities: severe side effects of modern targeted and immunotherapy of cancer and their management. *Crit Care*. 2017;21(1):89.

290. Mariotto AB, Yabroff KR, Shao Y, Feuer EJ, Brown ML. Projections of the cost of cancer care in the United States: 2010-2020. *J Natl Cancer Inst*. 2011;103(2):117-128.

291. Chan A-W, Hróbjartsson A, Haahr MT, Gøtzsche PC, Altman DG. Empirical evidence for selective reporting of outcomes in randomized trials: comparison of protocols to published articles. *JAMA*. 2004;291(20):2457-2465.

292. Wittes J. On changing a long-term clinical trial midstream. *Stat Med*. 2002;21(19):2789-2795.

293. Kirkham JJ, Dwan KM, Altman DG, et al. The impact of outcome reporting bias in randomised controlled trials on a cohort of systematic reviews. *BMJ*. 2010;340:c365.

294. Rankin J, Ross A, Baker J, O'Brien M, Scheckel C, Vassar M. Selective outcome reporting in obesity clinical trials: a cross-sectional review. *Clin Obes*. 2017;7(4):245-254.

295. Klara K, Kim J, Ross JS. Direct-to-Consumer Broadcast Advertisements for Pharmaceuticals: Off-Label Promotion and Adherence to FDA Guidelines. *J Gen Intern Med*. 2018;33(5):651-658.

296. Chen H, Cohen P, Chen S. How Big is a Big Odds Ratio? Interpreting the Magnitudes of Odds Ratios in Epidemiological Studies. *Communications in Statistics - Simulation and Computation*. 2010;39(4):860-864.

297. VanderWeele TJ, Ding P. Sensitivity Analysis in Observational Research: Introducing the E-Value. *Ann Intern Med*. 2017;167(4):268-274.

298. Dwan K, Altman DG, Clarke M, et al. Evidence for the selective reporting of analyses and discrepancies in clinical trials: a systematic review of cohort studies of clinical trials. *PLoS Med*. 2014;11(6):e1001666.

299. Als-Nielsen B, Chen W, Gluud C, Kjaergard LL. Association of funding and conclusions in randomized drug trials: a reflection of treatment effect or adverse events? *JAMA*. 2003;290(7):921-928.

300. Lundh A, Lexchin J, Mintzes B, Schroll JB, Bero L. Industry sponsorship and research outcome: systematic review with meta-analysis. *Intensive Care Med*. 2018;44(10):1603-1612.

301. Bassler D, Briel M, Montori VM, et al. Stopping randomized trials early for benefit and estimation of treatment effects: systematic review and meta-regression analysis. *JAMA*. 2010;303(12):1180-1187.

302. Food and Drug Administration. Presenting Quantitative Efficacy and Risk Information in Direct-to-Consumer Promotional Labeling and Advertisements: Guidance for Industry. Published October 2018. Accessed January 24, 2019. https://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/UCM623515.pdf

303. D'Agostino Sr RB, Massaro JM, Sullivan LM. Non-inferiority trials: design concepts and issues--the encounters of academic consultants in statistics. *Stat Med*. 2003;22(2):169-186.

304. Aberegg SK, Hersh AM, Samore MH. Empirical Consequences of Current Recommendations for the Design and Interpretation of Noninferiority Trials. *J Gen Intern Med*. 2018;33(1):88-96.

305. Lange S, Freitag G. Choice of delta: requirements and reality--results of a systematic review. *Biom J*. 2005;47(1):12-27; discussion 99-107.

306. Wangge G, Klungel OH, Roes KCB, de Boer A, Hoes AW, Knol MJ. Interpretation and inference in noninferiority randomized controlled trials in drug research. *Clin Pharmacol Ther*. 2010;88(3):420-423.

307. Wangge G, Klungel OH, Roes KCB, de Boer A, Hoes AW, Knol MJ. Room for improvement in conducting and reporting non-inferiority randomized controlled trials on drugs: a systematic review. *PLoS One*. 2010;5(10):e13550.

308. Garattini S, Bertele' V. Non-inferiority trials are unethical because they disregard patients' interests. *Lancet*. 2007;370(9602):1875-1877.

309. Piaggio G, Elbourne DR, Pocock SJ, Evans SJW, Altman DG, CONSORT Group. Reporting of noninferiority and equivalence randomized trials: extension of the CONSORT 2010 statement. *JAMA*. 2012;308(24):2594-2604.

310. Food and Drug Administration (FDA). Non-Inferiority Clinical Trials to Establish Effectiveness: Guidance for Industry. Published November 2016. Accessed March 12, 2019. https://www.fda.gov/downloads/Drugs/Guidances/UCM202140.pdf

311. Kudo M, Finn RS, Qin S, et al. Lenvatinib versus sorafenib in first-line treatment of patients with unresectable hepatocellular carcinoma: a randomised phase 3 non-inferiority trial. *Lancet*. 2018;391(10126):1163-1173.

312. Sanoff HK, Chang Y, Lund JL, O'Neil BH, Dusetzina SB. Sorafenib Effectiveness in Advanced Hepatocellular Carcinoma. *Oncologist*. 2016;21(9):1113-1120.

313. Rehal S, Morris TP, Fielding K, Carpenter JR, Phillips PPJ. Non-inferiority trials: are they inferior? A systematic review of reporting in major medical journals. *BMJ Open*. 2016;6(10):e012594.

314. Tsui M, Rehal S, Jairath V, Kahan BC. Most non-inferiority trials were not designed to preserve active comparator treatment effects. *J Clin Epidemiol*. Published online March 8, 2019. doi:10.1016/j.jclinepi.2019.03.003

315. Wayant C, Ross A, Vassar M. Noninferiority Trials - Oncology. Published March 20, 2019. Accessed May, 15, 2019. https://osf.io/mkg96/

316. Mesa RA, Kiladjian J-J, Catalano JV, et al. SIMPLIFY-1: A Phase III Randomized Trial of Momelotinib Versus Ruxolitinib in Janus Kinase Inhibitor–Naïve Patients With Myelofibrosis. *J Clin Orthod*. 2017;35(34):3844-3850.

317. D'Agostino RB Sr, Massaro JM, Sullivan LM. Non-inferiority trials: design concepts and issues - the encounters of academic consultants in statistics. *Stat Med*. 2003;22(2):169-186.

318. Fleming TR. Current issues in non-inferiority trials. *Stat Med*. 2008;27(3):317-332.

319. Munafò MR, Nosek BA, Bishop DVM, et al. A manifesto for reproducible science. *Nature Human Behaviour*. 2017;1(1):s41562-016 - 0021.

320. Collins FS, Tabak LA. Policy: NIH plans to enhance reproducibility. *Nature*. 2014;505(7485):612-613.

321. Rivoirard R, Bourmaud A, Oriol M, et al. Quality of reporting in oncology studies: A systematic analysis of literature reviews and prospects. *Crit Rev Oncol Hematol*. 2017;112:179-189.

322. Collaboration OS. Estimating the reproducibility of psychological science. *Science*. 2015;349(6251):aac4716.

323. Nosek BA, Errington TM. Reproducibility in Cancer Biology: Making sense of replications. *eLife Sciences*. 2017;6:e23383.

324. Lakens D, Page-Gould E, van Assen MA, et al. Examining the Reproducibility of Meta-Analyses in Psychology: A Preliminary Report. Published online 2017. https://osf.io/preprints/bitss/xfbjf

325. Jagsi R, Huang G, Griffith K, et al. Attitudes toward and use of cancer management guidelines in a national sample of medical oncologists and surgeons. *J Natl Compr Canc Netw*. 2014;12(2):204-212.

326. Cole Wayant, Matthew J Page, Matt Vassar. Reproducibility of Oncology Meta-Analyses. Published May 23, 2018. https://osf.io/kxj9z/

327. Moher D, Shamseer L, Clarke M, et al. Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015 statement. *Syst Rev*. 2015;4:1.

328. NCCN. NCCN Guidelines for Treatment of Cancer by Site. Accessed May 6, 2018. https://www.nccn.org/professionals/physician_gls/default.aspx#site

329. Higgins JPT, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. *BMJ*. 2003;327(7414):557-560.

330. DerSimonian R, Laird N. Meta-analysis in clinical trials. *Control Clin Trials*. 1986;7(3):177-188.

331. Page MJ, Moher D. Evaluations of the uptake and impact of the Preferred Reporting Items for Systematic reviews and Meta-Analyses (PRISMA) Statement and extensions: a scoping review. *Syst Rev*. 2017;6(1):263.

332. Vaughn K, Skinner M, Vaughn V, Wayant C, Vassar M. Methodological and reporting quality of systematic reviews referenced in the clinical practice guideline for pediatric high-blood pressure. *J Hypertens*. Published online July 24, 2018. doi:10.1097/HJH.0000000000001870

333. Scott J, Howard B, Sinnett P, et al. Variable methodological quality and use found in systematic reviews referenced in STEMI clinical practice guidelines. *Am J Emerg Med*. Published online June 14, 2017. doi:10.1016/j.ajem.2017.06.010

334. Ross A, Rankin J, Beaman J, et al. Methodological quality of systematic reviews referenced in clinical practice guidelines for the treatment of opioid use disorder. *PLoS One*. 2017;12(8):e0181927.

335. Peters JPM, Hooft L, Grolman W, Stegeman I. Reporting Quality of Systematic Reviews and Meta-Analyses of Otorhinolaryngologic Articles Based on the PRISMA Statement. *PLoS One*. 2015;10(8):e0136540.

336. Liu Y, Zhang R, Huang J, et al. Reporting quality of systematic reviews/meta-analyses of acupuncture. *PLoS One*. 2014;9(11):e113172.

337. Page MJ, Shamseer L, Altman DG, et al. Epidemiology and Reporting Characteristics of Systematic Reviews of Biomedical Research: A Cross-Sectional Study. *PLoS Med*. 2016;13(5):e1002028.

338. Higgins JPT GS, ed. Cochrane Handbook for Systematic Reviews of Interventions: online version (5.1.0, March 2011). *The Cochrane Collaboration*. Published online 2011. www.cochrane-handbook.org

339. Roundtree AK, Kallen MA, Lopez-Olivo MA, et al. Poor reporting of search strategy and conflict of interest in over 250 narrative and systematic reviews of two biologic agents in arthritis: a systematic review. *J Clin Epidemiol*. 2009;62(2):128-137.

340. Schünemann H, Brożek J, Guyatt G, Oxman A, ed. *GRADE Handbook for Grading Quality of Evidence and Strength of Recommendations*.; 2013.

341. Bailey CE, Hu C-Y, You YN, et al. Increasing disparities in the age-related incidences of colon and rectal cancers in the United States, 1975-2010. *JAMA Surg*. 2015;150(1):17-22.

342. Weinberg BA, Marshall JL, Salem ME. The Growing Challenge of Young Adults With Colorectal Cancer. *Oncology* . 2017;31(5):381-389.

343. National Comprehensive Cancer Network (NCCN). Colon Cancer. Published October 19, 2018. Accessed January 29, 2019. https://www.nccn.org/professionals/physician_gls/pdf/colon.pdf

344. National Comprehensive Cancer Network (NCCN). Rectal Cancer. Published August 7, 2018. Accessed January 29, 2019. https://www.nccn.org/professionals/physician_gls/pdf/rectal.pdf

345. Moher D, Liberati A, Tetzlaff J, Altman DG, PRISMA Group. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *Ann Intern Med*. 2009;151(4):264-269, W64.

346. Wayant C, Puljak L, Bibens M, Vassar M. Protocol: Risk of bias and reporting quality in systematic reviews underpinning colorectal cancer guidelines. Published April 29, 2019. Accessed April 29, 2019. https://osf.io/a24bu/.

347. Higgins J, Altman D, Sterne J. 8.3.3 Quality scales and Cochrane reviews. Cochrane Handbook for Systematic Reviews of Interventions [Version 5.1.0]. Published March 2011. Accessed April 18, 2019. https://handbook-5-1.cochrane.org/chapter_8/8_3_3_quality_scales_and_cochrane_reviews.htm

348. Propadalo I, Tranfic M, Vuka I, Barcot O, Pericic TP, Puljak L. In Cochrane reviews, risk of bias assessments for allocation concealment were frequently not in line with Cochrane's Handbook guidance. *J Clin Epidemiol*. 2019;106:10-17.

349. Saric F, Barcot O, Puljak L. Risk of bias assessments for selective reporting were inadequate in the majority of Cochrane reviews. *J Clin Epidemiol*. Published online April 19, 2019. doi:10.1016/j.jclinepi.2019.04.007

350. Babic A, Tokalic R, Amílcar Silva Cunha J, et al. Assessments of attrition bias in Cochrane systematic reviews are highly inconsistent and thus hindering trial comparability. *BMC Med Res Methodol*. 2019;19(1):76.

351. National Comprehensive Cancer Network. NCCN Categories of Evidence and Consensus. Accessed April 18, 2019. https://www.nccn.org/professionals/physician_gls/categories_of_consensus.aspx

352. Jiang J-B, Jiang K, Dai Y, et al. Laparoscopic Versus Open Surgery for Mid-Low Rectal Cancer: a Systematic Review and Meta-Analysis on Short- and Long-Term Outcomes. *J Gastrointest Surg*. 2015;19(8):1497-1512.

353. Zhao D, Li Y, Wang S, Huang Z. Laparoscopic versus open surgery for rectal cancer: a meta-analysis of 3-year follow-up outcomes. *Int J Colorectal Dis*. 2016;31(4):805-811.

354. Zhang F-W, Zhou Z-Y, Wang H-L, et al. Laparoscopic versus open surgery for rectal cancer: a systematic review and meta-analysis of randomized controlled trials. *Asian Pac J Cancer Prev*. 2014;15(22):9985-9996.

355. Xiong B, Ma L, Zhang C. Laparoscopic versus open total mesorectal excision for middle and low rectal cancer: a meta-analysis of results of randomized controlled trials. *J Laparoendosc Adv Surg Tech A*. 2012;22(7):674-684.

356. Vennix S, Pelzers L, Bouvy N, et al. Laparoscopic versus open total mesorectal excision for rectal cancer. *Cochrane Database Syst Rev*. 2014;(4):CD005200.

357. Arezzo A, Passera R, Scozzari G, Verra M, Morino M. Laparoscopy for rectal cancer reduces short-term mortality and morbidity: results of a systematic review and meta-analysis. *Surg Endosc*. 2013;27(5):1485-1502.

358. Trastulli S, Cirocchi R, Listorti C, et al. Laparoscopic vs open resection for rectal cancer: a meta-analysis of randomized clinical trials. *Colorectal Dis*. 2012;14(6):e277-e296.

359. Grimshaw J, Freemantle N, Wallace S, et al. Developing and implementing clinical practice guidelines. *Qual Health Care*. 1995;4(1):55-64.

360. Feder G, Eccles M, Grol R, Griffiths C, Grimshaw J. Clinical guidelines: using clinical guidelines. *BMJ*. 1999;318(7185):728-730.

361. Woolf SH, George JN. Evidence-based medicine. Interpreting studies and setting policy. *Hematol Oncol Clin North Am*. 2000;14(4):761-784.

362. Grilli R, Magrini N, Penna A, Mura G, Liberati A. Practice guidelines developed by specialty societies: the need for a critical appraisal. *Lancet*. 2000;355(9198):103-106.

363. Benson AB 3rd, Brown E. Role of NCCN in Integrating Cancer Clinical Practice Guidelines into the Healthcare Debate. *Am Health Drug Benefits*. 2008;1(1):28-33.

364. Wayant C, Cooper C, Turner D, Vassar M. Evaluation of the NCCN guidelines using the RIGHT Statement and AGREE II instrument: a cross-sectional review. *bioRxiv*. Published online 2018. doi:10.1101/442285

365. Hoffmann-Eßer W, Siering U, Neugebauer EAM, Brockhaus AC, Lampert U, Eikermann M. Guideline appraisal with AGREE II: Systematic review of the current evidence on how users handle the 2 overall assessments. *PLoS One*. 2017;12(3):e0174831.

366. Wayant C, Cooper C, Turner D ’arcy, Vassar M. RIGHT/AGREE 2 NCCN guidelines. Published October 19, 2018. Accessed October 19, 2018. https://osf.io/nh46r/

367. Uzeloto JS, Moseley AM, Elkins MR, et al. The quality of clinical practice guidelines for chronic respiratory diseases and the reliability of the AGREE II: an observational study. *Physiotherapy*. 2017;103(4):439-445.

368. Radwan M, Akbari Sari A, Rashidian A, Takian A, Abou-Dagga S, Elsous A. Appraising the methodological quality of the clinical practice guideline for diabetes mellitus using the AGREE II instrument: a methodological evaluation. *JRSM Open*. 2017;8(2):2054270416682673.

369. Larenas-Linnemann DES, Antolín-Amérigo D, Parisi C, et al. National clinical practice guidelines for allergen immunotherapy: An international assessment applying AGREE-II. *Allergy*. 2018;73(3):664-672.

370. Molino C de GRC, Leite-Santos NC, Gabriel FC, et al. Factors Associated With High-Quality Guidelines for the Pharmacologic Management of Chronic Diseases in Primary Care: A Systematic Review. *JAMA Intern Med*. Published online February 18, 2019. doi:10.1001/jamainternmed.2018.7529.

371. Yun X, Yaolong C, Zhao Z, et al. Using the RIGHT statement to evaluate the reporting quality of clinical practice guidelines in traditional Chinese medicine. *PLoS One*. 2018;13(11):e0207580.

372. Murad MH, Montori VM, Guyatt GH. Incorporating patient preferences in evidence-based medicine. *JAMA*. 2008;300(21):2483; author reply 2483-2484.

373. Zhang Y, Coello PA, Brożek J, et al. Using patient values and preferences to inform the importance of health outcomes in practice guideline development following the GRADE approach. *Health Qual Life Outcomes*. 2017;15(1):52.

374. Sackett DL, Rosenberg WMC, Muir Gray JA, Brian Haynes R, Scott Richardson W. Evidence based medicine: what it is and what it isn't. *BMJ*. 1996;312(7023):71-72.

375. Aslam S, Emmanuel P. Formulating a researchable question: A critical step for facilitating good clinical research. *Indian J Sex Transm Dis AIDS*. 2010;31(1):47-50.

376. Beitz JM. Writing the researchable question. *J Wound Ostomy Continence Nurs*. 2006;33(2):122-124.

377. Freedman RJ, Aziz N, Albanes D, et al. Weight and body composition changes during and after adjuvant chemotherapy in women with breast cancer. *J Clin Endocrinol Metab*. 2004;89(5):2248-2253.

378. Pardue SF, Fenton MV, Rounds LR. The social impact of cancer. *Dimens Oncol Nurs*. 1989;3(1):5-13.

379. Harms CA, Cohen L, Pooley JA, Chambers SK, Galvão DA, Newton RU. Quality of life and psychological distress in cancer survivors: The role of psycho-social resources for resilience. *Psychooncology*. 2019;28(2):271-277.

380. Nurgali K, Jagoe RT, Abalo R. Editorial: Adverse Effects of Cancer Chemotherapy: Anything New to Improve Tolerance and Reduce Sequelae? *Front Pharmacol*. 2018;9:245.

381. Franklin HR, Simonetti GP, Dubbelman AC, et al. Toxicity grading systems. A comparison between the WHO scoring system and the Common Toxicity Criteria when used for nausea and vomiting. *Ann Oncol*. 1994;5(2):113-117.

382. Young FE. The role of the FDA in the effort against AIDS. *Public Health Rep*. 1988;103(3):242-245.

383. Food and Drug Administration. Accelerated Approval. Published January 4, 2018. Accessed April 3, 2019. https://www.fda.gov/ForPatients/Approvals/Fast/ucm405447.htm

384. Prasad V, De Jesús K, Mailankody S. The high price of anticancer drugs: origins, implications, barriers, solutions. *Nat Rev Clin Oncol*. 2017;14(6):381-390.

385. Bishal Gyawali @oncology_bg. PFD (Progression free duration) is more accurate than PFS (Progression free survival). Twitter. Published December 12, 2019. Accessed February 25, 2021. https://twitter.com/oncology_bg/status/1205101042582638595

386. Poldervaart JM, Reitsma JB, Backus BE, et al. Effect of Using the HEART Score in Patients With Chest Pain in the Emergency Department: A Stepped-Wedge, Cluster Randomized Trial. *Ann Intern Med*. 2017;166(10):689-697.

387. Wiens BL, Zhao W. The role of intention to treat in analysis of noninferiority studies. *Clin Trials*. 2007;4(3):286-291.

388. Prasad V. Non-Inferiority Trials in Medicine: Practice Changing or a Self-Fulfilling Prophecy? *J Gen Intern Med*. 2018;33(1):3-5.

389. Egger M, Smith GD, Sterne JA. Uses and abuses of meta-analysis. *Clin Med* . 2001;1(6):478-484.

390. Goodman SN, Fanelli D, Ioannidis JPA. What does research reproducibility mean? *Sci Transl Med*. 2016;8(341):341ps12.

391. Hopewell S, Ravaud P, Baron G, Boutron I. Effect of editors' implementation of CONSORT guidelines on the reporting of abstracts in high impact medical journals: interrupted time series analysis. *BMJ*. 2012;344:e4178.

392. Dickersin K, Manheimer E, Wieland S, Robinson KA, Lefebvre C, McDonald S. Development of the Cochrane Collaboration's CENTRAL Register of controlled clinical trials. *Eval Health Prof*. 2002;25(1):38-64.

393. Miller J, Ross JS, Wilenzick M, Mello MM. Sharing of clinical trial data and results reporting practices among large pharmaceutical companies: cross sectional descriptive study and pilot of a tool to improve company practices. *BMJ*. 2019;366:l4217.

394. Salas-Vega S, Iliopoulos O, Mossialos E. Assessment of Overall Survival, Quality of Life, and Safety Benefits Associated With New Cancer Medicines. *JAMA Oncol*. 2017;3(3):382-390.

395. Davis C, Naci H, Gurpinar E, Poplavska E, Pinto A, Aggarwal A. Availability of evidence of benefits on overall survival and quality of life of cancer drugs approved by European Medicines Agency: retrospective cohort study of drug approvals 2009-13. *BMJ*. 2017;359:j4530.

396. Haslam A, Herrera-Perez D, Gill J, Prasad V. Patient Experience Captured by Quality-of-Life Measurement in Oncology Clinical Trials. *JAMA Netw Open*. 2020;3(3):e200363.

397. Covvey JR, Kamal KM, Gorse EE, et al. Barriers and facilitators to shared decision-making in oncology: a systematic review of the literature. *Support Care Cancer*. 2019;27(5):1613-1637.

398. Prasad V, Mailankody S. Research and Development Spending to Bring a Single Cancer Drug to Market and Revenues After Approval. *JAMA Intern Med*. 2017;177(11):1569-1575.

399. Perlis RH, Perlis CS. Physician Payments from Industry Are Associated with Greater Medicare Part D Prescribing Costs. *PLoS One*. 2016;11(5):e0155474.

400. Ahmed H, Turner S. Severe asthma in children-a review of definitions, epidemiology, and treatment options in 2019. *Pediatr Pulmonol*. 2019;54(6):778-787.

401. Stanton AL, Rowland JH, Ganz PA. Life after diagnosis and treatment of cancer in adulthood: contributions from psychosocial oncology research. *Am Psychol*. 2015;70(2):159-174.

402. Moher D, Naudet F, Cristea IA, Miedema F, Ioannidis JPA, Goodman SN. Assessing scientists for hiring, promotion, and tenure. *PLoS Biol*. 2018;16(3):e2004089.

# APPENDICES

*Table 16. Additional characteristics of included clinical trials.*

| **Characteristic** | | | **No. (%)** |
|---|---|---|---|
| *Journal* | New England Journal of Medicine | | 40 (41.7%) |
| | Lancet: Oncology | | 23 (24.0%) |
| | Journal of Clinical Oncology | | 12 (12.5%) |
| | Lancet | | 8 (8.3%) |
| | Blood | | 4 (4.2%) |
| | Leukemia | | 2 (2.1%) |
| | Lancet: Haematology | | 2 (2.1%) |
| | JAMA Oncology | | 2 (2.1%) |
| | Haematologica | | 2 (2.1%) |
| | European Journal of Cancer | | 1 (1.0%) |
| *Funding Source* | Industry | | 83 (86.5%) |
| | Mixed | Partial industry | 9 (9.4%) |
| | | No industry | 1 (1.0%) |
| | Public | | 2 (2.1%) |
| | Private | | 1 (1.0%) |
| *Group assignment* | Randomized | | 60 (62.5%) |
| | Single arm | | 32 (33.3%) |
| | Nonrandomized (2+ arms) | | 4 (4.2%) |
| *Blinding* | Double blind | | 28 (29.2%) |
| | Open label | | 68 (70.8%) |
| *Trial phase* | 1 | | 9 (9.4%) |
| | 1/2 (combined) | | 5 (5.2%) |
| | 2 | | 26 (27.1%) |
| | 3 | | 56 (58.3%) |
| *Hypothesis* | Superiority | | 57 (59.4%) |
| | Non-inferiority | | 3 (3.1%) |
| | N/A (single arm or non-comparative) | | 36 (37.5%) |

*Table 17.. General characteristics of systematic reviews underpinning the NCCN guidelines.*

| **Characteristic** | | **No. (%), All = 154** |
|---|---|---|
| *Year of Publication* | | |
| | 2011 | 11 (7.1%) |
| | 2012 | 31 (20.1%) |
| | 2013 | 29 (18.8%) |
| | 2014 | 22 (14.3%) |
| | 2015 | 41 (26.6%) |
| | 2016 | 15 (9.7%) |
| | 2017 | 4 (2.6%) |
| | 2018 | 1 (0.6%) |
| *Median number of meta-analyses* | | 14 (IQR 7.25 - 29.75) |
| *Any subgroup analyses reported* | | 96 (62.3%) |
| *Any sensitivity analyses reported* | | 66 (42.9%) |
| *Sources of Funding* | | |
| | Public (e.g., government) | 36 (23/.3%) |
| | Private (e.g., foundation) | 14 (9.1%) |
| | Hospital | 10 (6.5%) |
| | Industry | 5 (3.2%) |
| | None | 76 (49.4%) |
| | No statement | 23 (14.9%) |
| *Number of Cochrane reviews* | | 18 (11.7%) |
| *Number that adhered to PRISMA* | | 60 (39.0%) |

*Table 18. Characteristics of each index meta-analysis.*

| Characteristic | | All = 154 | PRISMA (n = 60) | non-PRISMA (n = 94) |
|---|---|---|---|---|
| *Type of outcome* | | | | |
| | All-cause mortality | 73 (47.4%) | 29 (48.3%) | 44 (46.8%) |
| | Comorbid event | 21 (13.6%) | 7 (11.7%) | 14 (14.9%) |
| | Recurrence of disease or therapy | 10 (6.5%) | 4 (6.7%) | 6 (6.4%) |
| | Disease response | 10 (6.5%) | 3 (5.0%) | 7 (7.4%) |
| | Cause-specific mortality | 8 (5.2%) | 4 (6.7%) | 4 (4.3%) |
| | Composite endpoint which includes mortality | 8 (5.2%) | 1 (1.7%) | 7 (7.4%) |
| | Length of hospital stay or operative time | 6 (3.9%) | 4 (6.7%) | 2 (2.1%) |
| | Patient-reported outcome | 6 (3.9%) | 0 (0.0%) | 6 (6.4%) |
| | Other physician assessed endpoint | 12 (7.8%) | 8 (5.2%) | 4 (4.3%) |
| *Described primary endpoint* | | 88 (57.1%) | 31 (51.7%) | 57 (60.6%) |
| *Median included studies* | | 8 (IQR 5 - 12) | 8 (IQR 5 - 11.25) | 7 (IQR 4 - 12) |
| *Median included patients* | | 1,914 (IQR 917 - 3,941) | 1896 (IQR 965 - 3883) | 1932 (IQR 891 - 3958) |
| *Effect measure* | | | | |
| | Hazard ratio | 58 (37.7%) | 23 (38.3%) | 35 (37.2%) |
| | Odds ratio | 40 (26.0%) | 14 (23.3%) | 26 (27.7%) |
| | Risk ratio | 35 (22.7%) | 14 (23.3%) | 21 (22.3%) |
| | Event rate | 9 (5.8%) | 3 (5.0%) | 6 (6.4%) |
| | Mean difference | 7 (4.5%) | 5 (8.3%) | 2 (2.1%) |
| | Response rate or median survival | 3 (1.9%) | 1 (1.7%) | 2 (2.1%) |
| | Standardized mean difference | 2 (1.3%) | 0 (0.0%) | 2 (2.1%) |
| *Random-effects model used* | | 91 (59.1%) | 43 (71.7%) | 48 (51.1%) |
| *Reported a subgroup analysis* | | 79 (51.3%) | 40 (66.7%) | 39 (41.5%) |
| *Reported a sensitivity analysis* | | 54 (35.1%) | 23 (38.3%) | 31 (33.0%) |

*Table 19. Risk of bias in all domains for individual studies.*

| Author | DOMAIN 1 Protocol & Eligibility Criteria | DOMAIN 2 Methods to identify and/or select studies | DOMAIN 3 Data collection & Appraisal | DOMAIN 4 Synthesis of findings | Consensus |
|---|---|---|---|---|---|
| *Ahmad, 2013* | Unclear | Unclear | Unclear | High | Unclear |
| *Amelung, 2015* | High | Unclear | High | Unclear | High |
| *Ardekani, 201* | Unclear | Unclear | Unclear | High | Unclear |
| *Arezzo, 2013* | Low | Unclear | Low | Low | Low |
| *Belum, 2013* | High | High | High | Unclear | High |
| *Berry, 2015* | Unclear | Unclear | Unclear | High | Unclear |
| *Böckelman, 2015* | Unclear | Unclear | Unclear | Unclear | Unclear |
| *Botrel, 2016* | Unclear | Unclear | Unclear | Unclear | Unclear |
| *Bujko, 2015* | High | High | Unclear | Unclear | High |
| *Chang, 2015* | High | High | Unclear | High | High |
| *Chung, 2011* | Low | High | High | Unclear | Unclear |
| *Ciliberto, 2012* | High | High | Unclear | High | High |
| *Dahabreh, 2011* | Low | High | Unclear | Unclear | Unclear |
| *Dai, 201* | High | High | Unclear | Unclear | High |
| *Di, 2013* | High | High | Low | High | High |
| *Elwood, 2016* | Low | Low | Unclear | Unclear | Unclear |
| *Filippo, 2015* | High | High | High | High | High |
| *Guo, 2016* | High | High | High | Unclear | High |
| *Hofheinz, 2016* | High | High | Unclear | High | High |
| *Huang, 2014* | Unclear | Unclear | Unclear | Unclear | Unclear |
| *Huang, 2014* | Unclear | Unclear | Unclear | Unclear | Unclear |
| *Jiang, 2015* | High | High | Low | Unclear | High |
| *Kidane, 2015* | Unclear | High | Unclear | Unclear | Unclear |
| *Lim, 2016* | High | Unclear | Unclear | High | High |
| *Liu, 2014* | High | High | Unclear | High | High |
| *Lu, 2015* | High | Unclear | Unclear | Unclear | Unclear |
| *Lv, 2013* | Unclear | Unclear | Unclear | Unclear | Unclear |
| *Macedo, 2012* | Unclear | Unclear | High | Unclear | Unclear |
| *Matsuda, 2015* | Unclear | Unclear | Unclear | High | Unclear |
| *Mirnezami, 2013* | High | Unclear | Unclear | Unclear | Unclear |
| *Ohtani, 2012* | Unclear | High | Unclear | Unclear | Unclear |
| *Petersen, 2012* | Unclear | Unclear | Unclear | Unclear | Unclear |

| | | | | | |
|---|---|---|---|---|---|
| *Petrelli, 2012a* | High | High | Unclear | Unclear | High |
| *Petrelli, 2012b* | Unclear | High | Unclear | Unclear | Unclear |
| *Petrelli, 2013* | High | High | Unclear | Unclear | High |
| *Pietrantonio, 2015* | High | Unclear | Unclear | Unclear | Unclear |
| *Pita-Fernández, 2015* | Unclear | Unclear | Unclear | Unclear | Unclear |
| *Qu, 2015* | Unclear | High | High | Unclear | High |
| *Rahbari, 2012* | Unclear | Unclear | Unclear | Unclear | Unclear |
| *Rahbari, 2013* | Unclear | Unclear | Unclear | High | Unclear |
| *Ranpura, 2011* | High | Unclear | High | Unclear | Unclear |
| *Rokkas, 2016* | High | High | Unclear | Unclear | High |
| *Rondelli, 201* | Unclear | Unclear | High | Unclear | Unclear |
| *Rowland, 2015* | Unclear | Unclear | Unclear | High | Unclear |
| *Sajid, 201* | Unclear | High | High | Unclear | High |
| *Schiphorst, 2015* | High | High | Unclear | Low | High |
| *Segelov, 2014* | Unclear | Low | Unclear | Low | Unclear |
| *Sorich, 2015* | High | Unclear | Unclear | Low | Unclear |
| *Theophilus, 2014* | Unclear | Unclear | Unclear | Unclear | Unclear |
| *Trastulli, 2012* | Unclear | Unclear | Low | Unclear | Unclear |
| *Vennix, 2014* | Low | Low | Unclear | Unclear | Low |
| *Wang, 2015* | High | High | Unclear | Unclear | High |
| *Wen, 2013* | Unclear | Unclear | Unclear | Unclear | Unclear |
| *Whitlock, 201* | Low | Low | Low | Unclear | Low |
| *Wu, 2012* | Unclear | High | High | Unclear | High |
| *Xiong, 2012* | Unclear | Unclear | Unclear | Unclear | Unclear |
| *Xu, 2017* | High | High | High | Unclear | High |
| *Zarak, 2015* | High | High | High | Unclear | High |
| *Zhang, 2012* | High | Unclear | High | Unclear | High |
| *Zhang, 2014* | Unclear | Unclear | Unclear | Unclear | Unclear |
| *Zhang, 2015* | Unclear | Unclear | Unclear | Unclear | Unclear |
| *Zhao, 2016* | High | Low | Unclear | Unclear | Unclear |

*Table 20. Adherence to PRISMA items for all individual studies.*

| **Study** | **Mention adherence to PRISMA** | **PRISMA Adherence (n/N, %)** | |
|---|---|---|---|
| *Ahmad, 2013* | Yes | 17/27 | 63.00% |
| *Amelung, 2015* | Yes | 23/27 | 85.20% |
| *Ardekani, 201* | Yes | 19/27 | 70.40% |
| *Arezzo, 2013* | Yes | 26/27 | 96.30% |
| *Belum, 2013* | No | 13/25 | 52.00% |
| *Berry, 2015* | No | 20/27 | 74.10% |
| *Böckelman, 2015* | No | 20/27 | 74.10% |
| *Botrel, 2016* | No | 23/27 | 85.20% |
| *Bujko, 2015* | No | 16/27 | 59.30% |
| *Chang, 2015* | No | 23/27 | 85.20% |
| *Chung, 2011* | No | 18/27 | 66.70% |
| *Ciliberto, 2012* | No | 16/25 | 64.00% |
| *Dahabreh, 2011* | No | 22/27 | 81.50% |
| *Dai, 201* | No | 20/27 | 74.10% |
| *Di, 2013* | No | 19/25 | 76.00% |
| *Elwood, 2016* | Yes | 22/27 | 81.50% |
| *Filippo, 2015* | Yes | 19/27 | 70.40% |
| *Guo, 2016* | No | 17/25 | 68.00% |
| *Hofheinz, 2016* | Yes | 17/27 | 63.00% |
| *Huang, 2014* | Yes | 19/27 | 70.40% |
| *Huang, 2014* | No | 16/25 | 64.00% |
| *Jiang, 2015* | Yes | 22/27 | 81.50% |
| *Kidane, 2015* | No | 23/27 | 85.20% |
| *Lim, 2016* | No | 22/27 | 81.50% |
| *Liu, 2014* | Yes | 21/27 | 77.80% |
| *Lu, 2015* | Yes | 19/27 | 70.40% |
| *Lv, 2013* | No | 21/27 | 77.80% |
| *Macedo, 2012* | No | 19/27 | 70.40% |
| *Matsuda, 2015* | Yes | 20/27 | 74.10% |
| *Mirnezami, 2013* | Yes | 20/27 | 74.10% |
| *Ohtani, 2012* | Yes | 16/25 | 64.00% |
| *Petersen, 2012* | Yes | 20/27 | 74.10% |
| *Petrelli, 2012a* | Yes | 23/27 | 85.20% |
| *Petrelli, 2012b* | No | 20/25 | 80.00% |
| *Petrelli, 2013* | No | 19/27 | 70.40% |

| *Petrelli, 2015* | **No** | **17/25** | **68.00%** |
|---|---|---|---|
| *Pietrantonio, 2015* | No | 19/27 | 70.40% |
| *Pita-Fernández, 2015* | Yes | 20/27 | 74.10% |
| *Qu, 2015* | No | 15/25 | 60.00% |
| *Rahbari, 2012* | Yes | 24/27 | 88.90% |
| *Rahbari, 2013* | Yes | 21/27 | 77.80% |
| *Ranpura, 2011* | No | 21/27 | 77.80% |
| *Rokkas, 2016* | Yes | 21/27 | 77.80% |
| *Rondelli, 201* | Yes | 22/27 | 81.50% |
| *Rowland, 2015* | No | 20/27 | 74.10% |
| *Sajid, 201* | No | 20/27 | 74.10% |
| *Schiphorst, 2015* | No | 21/27 | 77.80% |
| *Segelov, 2014* | Yes | 18/27 | 66.70% |
| *Sorich, 2015* | No | 20/27 | 74.10% |
| *Theophilus, 2014* | No | 20/27 | 74.10% |
| *Trastulli, 2012* | No | 20/27 | 74.10% |
| *Vennix, 2014* | No | 22/27 | 81.50% |
| *Wang, 2015* | Yes | 23/27 | 85.20% |
| *Wen, 2013* | No | 21/27 | 77.80% |
| *Whitlock, 201* | No | 17/27 | 63.00% |
| *Wu, 2012* | No | 19/27 | 70.40% |
| *Xiong, 2012* | No | 20/25 | 80.00% |
| *Xu, 2017* | Yes | 22/25 | 88.00% |
| *Zarak, 2015* | No | 16/27 | 59.30% |
| *Zhang, 2012* | No | 16/25 | 64.00% |
| *Zhang, 2014* | No | 17/27 | 63.00% |
| *Zhang, 2015* | No | 21/27 | 77.80% |
| *Zhao, 2016* | No | 19/27 | 70.40% |

VITA

Christian Cole Wayant

Candidate for the Degree of

Doctor of Philosophy

**Dissertation**: RIGOR AND REPRODUCIBILITY OF CANCER MEDICINE EVIDENCE

**Major Field**: Biomedical Sciences

**Biographical**:

*Education*:

Completed the requirements for the Doctor of Philosophy in Biomedical Sciences at Oklahoma State University, Stillwater, Oklahoma in May, 2021.

Completed the requirements for the Bachelor of Science in Biology at the University of Oklahoma, Norman, Oklahoma in 2015.

*Experience*:

Awarded Ruth L. Kirschstein National Research Service Award Individual Fellowship for Students at Institutions Without NIH-Funded Institutional Predoctoral Dual-Degree Training Programs (Parent F30)

Awarded AHRQ Health Services Research Dissertation Program (R36)

Inaugural Doug Altman Scholar 2019, OSU-CHS Student Doctor of the Year 2019, OSU Graduate College Distinguished Graduate Fellow, Women's Faculty Council Outstanding Student Researcher, COSGP National Student Researcher of the Year (Top 10 Finalist)

Invited, oral presentations at EBM Live (Oxford, England, UK; 2019), Peer Review Congress (Chicago, IL; 2017), Society for Clinical Trials (New Orleans, LA; 2019 | Montreal, Quebec, CA; 2016).