6-29-2017 12:00 AM

Evidence Reversal: An exploratory analysis of randomized controlled trials from the New England Journal of Medicine

Riaz G. Qureshi
The University of Western Ontario

Supervisor
Dr. Janet Martin
The University of Western Ontario

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

ABSTRACT

BACKGROUND: Evidence Reversal (ER) is the phenomenon whereby new and stronger evidence contradicts previously established evidence.

OBJECTIVES: To quantify evidence reversals and to determine characteristics associated with reversibility.

METHODS: Original articles from the New England Journal of Medicine (2000 to 2016) were screened for three inclusion criteria: tested a clinical practice; Randomized Controlled Trial design; and tested an established clinical practice. The proportion of RCTs that represented ER was determined. Association of trial characteristics with reversal was explored using logistic regression in order to inform a potential framework of reversibility.

RESULTS: In total, 611 RCTs met the inclusion criteria, of which 54% were evidence reversals. Based on variables associated with ER, a reversibility framework was proposed, comprised of eight trial characteristics.

CONCLUSION: More than 50% of RCTs published in the NEJM that test established practices are evidence reversals. The characteristics of RCTs that are associated with reversal will inform future research to further understand reversibility.

KEYWORDS: Evidence Reversal, Medical Reversal, Evidence-Based Medicine, Evidence Synthesis, Randomized Controlled Trials, Adoption, De-Adoption, Implementation, De-Implementation, Decision-Making

ACKNOWLEDGEMENTS

This thesis is dedicated to my parents: Arif and Susan Qureshi – for directing me to choose my own fate, instilling in me a need to do my best, and supporting me in everything I have done. I wouldn't be anywhere without their guidance and all they have given me.

Special thanks to my girlfriend, Stephanie Smith, who has been through much with me and has always supported my studies, knowing how important they are to me.

Special thanks also to Dr. Martin for providing me with an amazing opportunity and being an incredible mentor – one who was always enthusiastic about what I had going on and every suggestion and idea that I had – and of course to Desirée Sutton, without whom much of this thesis would not have been possible. The amount of work that she initiated before passing the mantle onto me and the work that she continued to contribute to this thesis is truly staggering and I am constantly thankful for her help in making this project a reality within the given time frame.

Thanks also to the members of my supervisory committee: Dr. Neil Klar, Dr. Davy Cheng, and Dr. Philip Jones. All provided expertise in areas where I was uncertain and asked questions that made me approach perspectives and problems that I had not yet considered.

I would also like to acknowledge Jessica Moodie for all of her help in organizing meetings between my committee members and myself, and in collecting articles that I could not find.

And lastly, but certainly not least, I would like to thank all of my wonderful friends that I have made in this program for the many insightful discussions about life, our future careers, class assignments, and even directions with our theses themselves. In particular: Dr. Lenny Guizzetti, Alex Ratzki-Leewing, Josh Cerasuolo, and Patrick Kim. I have made some friendships that will burgeon as we move from school into our respective careers and I look forward to working with all of them in the future.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF APPENDICES

LIST OF ABBREVIATIONS

BMJ	British Medical Journal
CI	Confidence Interval
CM-A	Cumulative Meta-Analysis
EBM	Evidence Based Medicine
EBP	Evidence Based Practice
ER	Evidence Reversal
GRADE	Grading of Recommendations Assessment, Development and Evaluation
HL	Hosmer-Lemeshow
ICER	Incremental Cost-Effectiveness Ratio
JAMA	Journal of the American Medical Association
KT	Knowledge Translation
M-A	Meta-Analysis
MCID	Minimally Clinically Important Difference
MR	Medical Reversal
NEJM	New England Journal of Medicine
NICE	National Institute for Health and Care Excellence
PICOTS	Population Intervention Comparison Outcome Timing Study design / Setting
PRISMA	Preferred Reporting Items for Systematic Reviews and Meta-Analyses
RCT	Randomized Controlled Trial
ROB	Risk of Bias
SR	Systematic Review

LIST OF TERMS

Clinical Practice Guideline: Statements that have been systematically developed to assist practitioner and patient decisions about appropriate health care for specific clinical circumstances

Disinvestment: The processes by which health care systems or services partially or completely withdraw resources from any existing health care practices, procedures, technologies, or pharmaceuticals that are deemed inefficient for health resource allocation (due to low health gain for their cost).

De-implementation: The cessation of practices that are not evidence-based (aka: de-adoption).

Evidence Based Medicine: The judicious and conscientious use of the best available evidence, clinical expertise, and patient values, in determining the best course of action for treating a patient in a clinical setting.

Evidence Replacement: A new medical practice – tested in a properly designed trial – surpassing and replacing an older standard of care.

Evidence Equivalent or Less: A new medical practice – tested in a properly designed trial – failing to surpass an older standard of care (aka: "Back to the drawing board").

Evidence Confirmation: An existing medical practice – tested in a properly designed trial – being found superior to a lesser or prior standard (aka: "Reaffirmation").

Medical Reversal: The phenomenon whereby new studies – better powered, controlled, or designed than their predecessors – contradict a current practice, leading to subsequent de-adoption.

Evidence Reversal: The phenomenon whereby new evidence that is stronger than its predecessors contradicts previously established evidence.

Evidence Maturity: A stage of research at which the culmination of evidence for a practice allows a decision to be made regarding its effectiveness and associated harms and benefits in clinical practice.

Evidence Stability: An aspect of cumulative meta-analysis that describes the shifts over time in the accruing evidence about whether an intervention works. Derived from the flow or trend in the running estimate of effect generated over sequential meta-analyses.

Evidence Sufficiency: An aspect of cumulative meta-analysis that describes whether the meta-analytic database adequately demonstrates whether an intervention works. Derived from the number of hypothesis tests that have been conducted for the intervention of interest.

Fragility Index: The minimum number of patients whose status would be required to change from a non-event to an event to turn a statistically significant result into a non-significant result. Similarly applied, 'Reverse Fragility Index' is the minimum number of patients whose status would have to change to turn a non-significant result into a significant result. Can only be applied to trials with dichotomous outcomes and 1:1 randomization. Smaller numbers indicate a more "fragile" trial result.

GRADE: 'Grading of Recommendations Assessment, Development and Evaluation' is an approach to grading the quality or certainty of evidence and strength of recommendations to guide clinical practice and the development of standardized guidelines.

Hard (clinical / patient-important) Outcome: Study endpoints for intervention effect that are definite and clinically meaningful. Demonstrate the effect, or lack thereof, with outcomes that patients care about (i.e. the way the patient feels, functions, or survives). Examples: all-cause mortality, myocardial infarction, stroke, validated Quality of Life.

Immature Evidence: Stages of research wherein an insufficient amount of evidence has accumulated to inform a decision regarding the effectiveness and related harms and benefits of a practice, thereby necessitating further study.

Knowledge Translation: The process of synthesizing, disseminating, and applying knowledge to improve health, provide more effective health services and products, and strengthen the health care system.

Meta-Analysis: The pooling of data from multiple clinical studies that address the same topic to assemble a database that is large enough to statistically reach a conclusion regarding the practice in question.

Realist Review: A strategy that is focused on explanation as opposed to judgement in the synthesis of research about a topic. Instead of demonstrating whether a practice or program is effective, they seek to describe the mechanism of how they work in particular contexts and settings.

Surrogate Outcome: Study endpoints for intervention effect that are not clinically meaningful. More easily and quickly ascertained than hard clinical outcomes, but not always correlated appropriately with patient-important outcomes that are relevant to the disease or intervention being studied. Examples: change in blood pressure, progression-free survival, levels of blood biomarkers.

Systematic Review: An approach to assessing the evidence base and providing an unbiased judgement on a question that is systematic, transparent, and reproducible.

CHAPTER 1

Introduction to Evidence Reversal

Riaz Qureshi

Note: After this chapter of this thesis was drafted, a manuscript to introduce concepts, contained within this chapter was drafted for publication. As such, there are many similarities in the sections of the paper and this chapter. However, I was able to expand this chapter and go into more detail than the paper due to space and formatting restrictions within the publication.

Citation: **Qureshi R**, Sutton D, & Martin J. (2016). Approaching Evidence Reversal and Medical Reversal – When to say, "Enough is enough." Ready for submission for publication to Mayo Clinic Proceedings, Apr 2017.

Chapter summary: This chapter presents an introduction to the phenomenon of reversal and describes some of the issues that surround research in this field, as well as the current tools in the field of reversal research and the proposed rationale and impact that this research will have on future research.

CHAPTER 1

1.0 WHAT ARE MEDICAL REVERSAL AND EVIDENCE REVERSAL?

A medical reversal is the phenomenon whereby a medical practice, procedure, or technology that has been embraced by the medical community loses its standing when better conducted studies show that it is not as effective as was thought, or even harmful to the population.[1–3] As medicine has advanced, so too has study methodology, and new trials are often superior – through better design, greater power, or more appropriate control groups – to the original studies of a medical practice.[1,2] When such superior studies are conducted to test the effectiveness of current clinical practices, they may contradict the original studies – with results that are incongruent with the beliefs and practices of the general medical community – and find that the current practice, procedure, or technology, is inferior to a prior standard of care, does not produce the promised results, or is even more harmful than beneficial to a patient's health.[2]

Evidence reversal occurs when new evidence comes to light and shows that the established evidence, often based on a combination of low quality research or limited availability, is incorrect.[1] While similar to the phenomenon of 'medical reversal,' the phenomenon of 'evidence reversal' differs in two regards. Firstly, the term 'medical reversal' limits the phenomenon to clinical practices, whereas 'evidence reversal' expands this definition beyond medicine alone to also include other fields relevant to health and healthcare, such as public and population health. Secondly, 'medical reversals' refers only to practices that are already adopted and implies de-adoption after reversal, whereas 'evidence reversal' refers only to the evidence for the practices as being reversed and the practices, if already adopted, may continue. It is for this reason that we propose

'evidence reversal' as being a more appropriate term to describe contradictory findings of new studies, and consequently, more appropriate than 'medical reversal.'

While evidence reversal may arise from the findings when a trial investigates an already established intervention, there are several other possible outcomes when studying interventions in general. When an established intervention is tested and found to be inferior to what was originally believed, this is a reversal, as has previously been established.[4] As the counterpart to reversal: when an established intervention is tested by newer studies and found to be as good as, or better than, was originally believed, this is termed "evidence reaffirmation."[1,4] When a new practice, device, or intervention is being tested and is found to be superior to an old standard of care, this constitutes an "evidence replacement."[1,4] The counterpart to evidence replacement is sometimes referred to as "back to the drawing board" – which occurs when a trial that tests a new practice, device, or intervention finds it equal to or no better than an old standard of care.[1,4]

The term 'medical reversal'– as applied to describing the phenomenon of new trials contradicting clinical practice – was first used in 2011 by Prasad and colleagues.[1] While the designation is still in its infancy and the term is not well known, physicians are familiar with the phenomenon of evidence reversal and subsequent de-implementation of established clinical practices.[5] Awareness of the phenomenon is increasing, even within the general public, as examples of reversals have been highlighted in the medical news of the popular media.[6,7]

1.1 FREQUENCY OF MEDICAL REVERSAL

Given the frequency with which guidelines and practices change in the medical literature, several researchers have tried to quantify the rate of reversal. Ioannidis and

colleagues found that among highly influential studies published in the New England Journal of Medicine (NEJM), Journal of the American Medical Association (JAMA), and Lancet between 1990 and 2003 that have been cited more than 1000 times, 16% were found to be contradicted by subsequent studies and an additional 16% were found to have smaller effects than initially found.[8] In further assessments of all original research articles published in the NEJM, Prasad and colleagues suggest that reversal could be even more prevalent. The proportion of trials published in 2009 that tested an established medical practice and found contradictory evidence constituting a reversal was 46%.[1] A similar assessment over a 10-year period, from 2001 to 2010, found that 40.2% of trials found contradictory evidence for their tested practice.[4]

1.2 EXAMPLES OF EVIDENCE AND MEDICAL REVERSAL

Three debated reversals include: stenting for stable coronary artery disease, vertebroplasty for osteoporotic fracture, and hormone replacement therapy for prevention of coronary heart disease.

1.2.1 Percutaneous coronary intervention for stable coronary artery disease

Ever since their invention, coronary artery stents – a small wire mesh tube designed to expand and open an artery with stenosis – have been used to treat people with myocardial infarction (MI). Placing a stent in an occluded artery at the site of blockage – percutaneous coronary intervention (PCI) – opens the vessel and restores blood flow, improving the chances of surviving the event. Due to their effectiveness in restoring flow to a damaged artery, PCI was also used in the treatment of typical angina – recurring chest pain with exertion that is experienced as a direct result of coronary artery disease (CAD). The physiology and mechanism for action was logical, and patients reported

feeling better after undergoing the procedure. However, a 2007 randomized, blinded trial of PCI plus medical therapy versus medical therapy alone in patients with stable CAD found that it did not reduce the risk of death (both cardiac and all-cause mortality), recurrent MI, stroke, hospitalisation for acute coronary stenosis, or revascularization.[9] This trial showed that while PCI for stable CAD relieved some symptoms, such as typical angina, for a brief time, placing a stent did nothing to improve patient survival or risk of future cardiovascular events and was therefore not as effective as was believed while subjecting patients to the risks associated with surgical intervention, such as: anaesthesia, infection of site wound, and hospital stay.[9] Despite the findings of this trial, the practice persists and is consequently more appropriately described as an evidence reversal than a medical reversal.

1.2.2 Vertebroplasty for osteoporotic fracture

In the early 1990s, a simple outpatient procedure for the management of osteoporotic spinal fractures became popular and gained widespread use. Vertebroplasty involves the injection of medical cement into fractured spinal bone with the intention of restoring original shape, stabilizing the fragments, and reducing pain from the fracture.[10] The procedure appeared to work: patients who underwent the procedure experienced drastically reduced pain and disability.[10] Based on these reports and several early trials that did not include controls, vertebroplasty was added to the list of Medicare-funded procedures in 2001.[11] Vertebroplasty quickly became a multi-million dollar industry and the number of procedures performed each year increased from 14,142 in 2001, to 29,090 in 2005.[11] However, the evidence for this procedure was reversed in 2009 when two randomized and double-blinded trials of vertebroplasty versus a sham procedure (simple salt water injection) found no difference between groups in response to treatment:

saltwater injection caused just as much reduction in pain and disability as medical cement.[12,13] Although the two trials provided convincing evidence to discredit the practice, its use has persisted and is consequently representative of evidence reversal.

1.2.3 Hormone replacement therapy for prevention of coronary heart disease

Prescription of hormone replacement therapy (HRT) – a combination of estrogen and/or progestin – to reduce the symptoms of menopause and as primary prevention for coronary heart disease (CHD) has been routine since the mid-1960s.[14–17] These endogenous hormones are critical for a number of physiologic processes including the reduction of osteoporotic bone loss, cardiovascular health, reproductive function, and temperature regulation dependent on hormonal homeostasis.[18] Following endogenous estrogen reduction after menopause, women experience increased risk of osteoporosis and bone fracture, MI, stroke, uterine and vaginal wall changes, and hot flushes.[18,19] Treatment with exogenous estrogen and progestin was a physiologically sound solution to improve bone mineral density and reduce the risk of MI, stroke, and other perimenopausal symptoms. The therapy seemed to work for select symptoms such as hot flushes and bone mineral density. However, two randomized controlled trials conducted between in 1993 and 1998 reported that women receiving HRT were at significantly higher risks for CHD, stroke, breast cancer, pulmonary embolism, and venous thromboembolic events than women not receiving HRT.[15,20] The risk-to-benefit ratio of HRT was too great for a primary prevention of CHD and osteoporosis, and the therapy quickly fell out of favour among post-menopausal women, although its use in other population subgroups remains contested.[14] While HRT use continues for some women,

the contraindication and cessation of routine use in most post-menopausal women is what designates it as an example of a medical reversal.

1.3 WHY SHOULD WE CARE ABOUT REVERSALS?

Reversals of evidence, whether in medicine or other health-related disciplines are an important phenomenon to the scientific community and the population as a whole. Society should care about this phenomenon because reversals pose several real dangers to clinical practice if left unchecked. Minimizing the impact and occurrence of reversals caused by the premature adoption of practices would benefit society in several ways.

1.3.1 The dangers of unnecessary reversal

Although the phenomenon of evidence reversal is a natural consequence of the scientific method – contradicting prior beliefs when new information and better methods to test those beliefs are available – the premature implementation of practices (whether they be medical, public health, or population based) that may have little to no benefit, or are potentially harmful can lead to serious consequences for society. These risks of premature adoption are often discussed in the literature as being the harms associated with reversal.

One of the primary dangers of reversal is unnecessary cost (i.e. wasted resources). Any technology or practice that does not work as intended, especially medical, has no place in the market. However, new technologies are often promoted by industry without a complete understanding of their effectiveness – often for uses for which they have not been tested nor approved. Furthermore, it is not only industry that does this as governments will support technologies or interventions that have a demonstrated need in their population, based on whatever evidence is available at the time.[21] The presence of

reversed practices in medicine places an unnecessary burden on the healthcare system as the government utilizes limited resources to provide services that are no better, or worse than, previously implemented standards of care, placebo, or even no intervention.[2,3,22]

Another danger of reversals is the potential risk at which those who receive reversed interventions have been placed for no benefit. Medical practice revolves around the principles of beneficence and non-maleficence: doing what is best for one's patient and not causing undue harm. These must always be kept in balance when determining the most suitable intervention: what level of risk is acceptable, given the expected benefit that the patient should receive? When patients receive reversed medical practices, they have undergone unnecessary risk for less benefit than they believed they were receiving, and this is ethically and morally wrong.

Another danger of reversal in medicine is the undermining of trust in the medical system. It is generally accepted that the public trusts that physicians know what treatment is best for their health problems and will administer a suitable therapy that has their interests in mind.[23,24] The core tenet of evidence-based medicine (EBM) is the integration of clinical expertise, best available evidence, and patient preferences in choosing the most suitable intervention.[25] However, all clinicians know that there are times when no clear path is available and sometimes the best that can be done is an educated guess. When patients receive multiple misdiagnoses or mistreatments they may lose faith in the medical system. In addition, the media portrayal of medical uncertainty, changing guidelines, and exaggerated claims of the benefits or risks of practices compound this mistrust. This damage that is done to a patient's or clinician's faith in the medical system

may be irreparable and undermine the ability of the system to help these people in the future.[26]

A further danger of medical and evidence reversals lies in the difficulty associated with removing an already established standard from the scientific community: de-adoption.[27] It is an established fact of knowledge translation that it takes many years for practices to be implemented in clinical care or for a scientific technique to be adopted. However, it is more difficult to remove an engrained intervention, technology, or paradigm from practice because the scientific community is not unbiased.[28,29] Practitioners will often justify the continued use of a popular standard or practice, despite evidence that it does not work as was originally believed. Once something has been reversed, there is no guarantee that it will cease to be used.

1.3.2 The benefits of eliminating unnecessary reversal

Premature uptake of practices, before sufficient evidence exists (leading to subsequent unnecessary evidence reversals when better evidence accrues), is common in medicine and poses a real risk to the health of the population. If the incidence of reversed practices could be reduced, or the impact of the phenomenon minimized, all members of society would benefit. The benefits of reducing the amount of medically reversed practices and technologies are all complementary to the dangers that accompany reversal, as discussed above.

Eliminating or reducing unnecessary evidence reversals due to premature knowledge translation would improve the overall health of the population because potentially harmful or ineffective practices may be stopped earlier or prevented from implementation. Many medical practices that are reversed place the recipients at risk for little to no benefit. It logically follows that reducing the number of practices that do not

work as planned or are more harmful than originally believed would subsequently reduce the burden of harm that is placed on the population.

If the number of reversals due to premature adoption were reduced, there would be increased trust in the medical community. Fewer practices would be implemented that later need to be de-adopted. This may lead patients to put more trust into their physicians and to be more open to seeking medical care when it is necessary.

With a reduction in reversals due to premature adoption, government administrators and policy makers may see an increase in available funding. This increase may be possible if premature conclusions based on insufficient evidence could be minimized, to prevent the premature uptake of practices. This prevention would lead to a reduction in unnecessary expenditure as money that was previously wasted on technologies and interventions that are no better or worse than placebo would be available for use in areas based on adequate evidence and proven efficacy.

1.4 APPROACHING EVIDENCE REVERSAL

The phenomenon of evidence reversal is difficult to approach because of its inherent ethical and logistic challenges. The primary dangers of reversal revolve around the fact that practices and paradigms that need to be reversed are often already engrained in the scientific community and/or widely believed. In approaching the phenomenon of evidence reversal – and in consideration of these dangers – it is vital to consider where the burden of proof may lie for identifying reversals.

Given the difficulties surrounding de-adoption of established practices, the most effective approach to minimizing their impact is to stop them before they gain a strong presence in clinical practice and population health. However, one of the difficulties in

identifying reversals before they have been adopted is establishing where the responsibility for identification lies.[22,26] Using the arguments presented by Prasad and Cifu for the identification of medical reversals, we believe that the burden of proof lies primarily with manufacturers, researchers, and regulatory bodies who develop and approve interventions and practices to ensure effectiveness before implementation.[26]

In asking this question, it is easiest to rule out where the burden of proof does not lie, and this is with the patients. Undoubtedly, the general population that is served by the medical community suffers when technologies that will be later reversed see widespread dissemination and use. As has been outlined in previous sections, reversals have many dangers and the public has the right to interventions that have proven efficacy.[22]

The burden of proof for identifying reversals must lie partly with physicians as the administrators of therapies. However, beyond their physiological knowledge and personal clinical experience and expertise, they can only know as much about the interventions that they prescribe as is given to them by the researchers and industry. The burden of proof lies in part with physicians, as they must be cognizant of the evidence for a treatment's efficacy before prescribing it to their patients. It is a physician's ethical duty to act in the best interests of the patient and when technologies are put into practice before their effects are fully understood, physicians take a risk in their prescription as they may be unknowingly putting the patient at increased risk of unnecessary harm.[22]

Following the description of the dangers of medical reversal, the rationale for why the burden of proof lies primarily with the governmental institutions that support the research and grant regulatory approval, the industry that creates the interventions, and the researchers who study their effectiveness and efficacy in clinical practice is clear: 1) it is

very difficult and costly to remove an already established practice from the field and it is safer and more efficient to confidently determine an intervention's efficacy and effectiveness before it is implemented than after it has been implemented – which is the responsibility of governmental institutions that support research and grant regulatory approval; 2) the actual proportion of interventions that have clinically relevant and significant impacts on patient important and meaningful outcomes is very low – which is the responsibility of the industry that creates interventions; and 3) the implementation of practices that may later be reversed is a waste of valuable health-care resources that could be avoided by adequately studying their effects before promoting their use – which is the responsibility of the researchers who study effectiveness and efficacy in practice and the agencies that fund research.[22]

1.5 TOOLS FOR APPROACHING EVIDENCE REVERSAL

In knowing where the burden of proof lies, consideration must be given to the methods and tools that are currently employed for identifying and reducing the impact of evidence reversals. While standards of practice exist in all scientific disciplines, new findings will always require knowledge dissemination before they can be implemented. As such, there currently exist several tools that attempt to mitigate the effects of unnecessary reversals by providing evidence-based recommendations: clinical guidelines, knowledge translation, and various tools for de-implementation.

An ideal clinical guideline should serve to inform practitioners and patients of the most appropriate treatment or course of action in any given circumstance. While there are many faults with the current processes employed in creating clinical guidelines – faults which themselves may sometimes lead to premature adoption and unnecessary reversal –

an ideal clinical guideline that is rigorous, unbiased, and uses the best available evidence should provide a recommendation for the most appropriate care within the context of the quality and quantity of available evidence.

Similarly, an ideal translation of research findings into practice could lead to a reduction in premature uptake of practices because clinicians and policy makers would consequently know which practices have a proven efficacy and which do not. Despite this, however, there remain a plethora of potential reasons as to why a practice may be prematurely adopted or remain in use after it has been reversed.

In attempting to reduce the impacts of unnecessary reversals and premature adoption, many different campaigns and programs have been developed to aid in the de-adoption of reversed practices and increase awareness of the value of different practices. These programs attempt to summarize the totality of evidence and provide recommendations to practitioners and even the general public to inform better health care.

In providing these recommendations for the implementation of new practices or de-adoption of established practices, an important consideration is the maturity of the evidence base to support the practice. Sufficiency and stability are characteristics to describe accumulated evidence and provide a measure of evidence maturity: the point at which an intervention has been studied enough that conducting another test no longer provides any information of value. There are several different methods of assessing the sufficiency and stability of evidence to aid decision-making including: cumulative meta-analysis, trial sequential analysis / monitoring, Bayesian analysis, value of information analysis, GRADE (Grades of Recommendation Assessment, Development, and Education), and the fragility index. Chapter 7 provides a detailed discussion of the

strengths and weaknesses of these tools and how the results of this thesis may be applied to inform their development and use in the context of medical and evidence reversal.

1.6 CAUSES OF EVIDENCE REVERSAL

Evidence reversal is a complex phenomenon that occurs when new research contradicts the established evidence for a claim or belief, suggesting it is not what was originally believed. Evidence reversal has many causes. The causes of evidence reversal are related to the characteristics of the original research itself – including characteristics of the innovation being studied – that played a role in the misguided investment, dissemination, and utilization of the practice, procedure, or technology that must consequently be reversed.

One of the common causes of reversal has already been shown in previous examples is a strong belief in the pathophysiological model that leads to the assumption that intervening on parts of the pathophysiologic causal pathway will translate to effectiveness, despite never demonstrating either effectiveness or efficacy with respect to clinically meaningful outcomes in a trial setting. Placing a stent in an artery with severe stenosis when someone is experiencing chest pain should prevent cardiac-related mortality or MI, but it does not. Neither does injecting medical cement into fractured bone to stabilize the fracture have an effect beyond a simple saline solution, even though it theoretically should. These seemingly logical pathways that are common to many evidence reversals are important to keep in mind because they demonstrate that unless there is direct evidence of an effect on an important outcome of interest – particularly clinically relevant outcomes – it is difficult to know the net effect a proposed intervention will have on a population in practice.[22]

There are many clinical practices that are established based on tradition and have never truly been tested in a randomized trial.[30] Such practices, sometimes called "sacred cows," are often based on positive observed effects within a pathophysiological model that do not translate to meaningful clinical outcomes.[31,32] Examples of practices that were used without proven efficacy, until they were reversed by trials, include: non-invasive measurement of blood pressure in children, oxygen administration for patients with chronic obstructive pulmonary disorder, and supplemental oxygen administration for acute MI.[33] A further complexity that these "sacred cows" impose on health care is that such practices are not easily tested to find their true value. Practices that have been used for a long time and are engrained cannot easily have their efficacy assessed because they are overwhelmingly believed to be effective, thus failing to satisfy the principle of clinical equipoise that is necessary to ethically justify randomization to not receive the therapy, and therefore would be seen as unethical by many practitioners.[34]

Trusting a physiological model may also lead to a related cause of evidence reversal, which is overgeneralization to non-study populations.[35] The expanded application of interventions to populations for which they have never been tested is a common cause of reversal.[1] Some practices may only be reversed for a particular indication (e.g. PCI is effective at saving the lives of those with MI, but has been reversed for preventing future heart-related incidents among those with stable CAD).[1,9]

Over reliance on physiological models is also directly related to another cause of medical reversals: the use of surrogate outcomes in trials that do not appropriately represent important clinical outcomes.[22,36] Surrogate outcomes (e.g. blood pressure, bone mineral density, tumour growth) are often used as endpoints in studies of new

interventions because they are cheaper and require less time to get results than using clinically meaningful outcomes (e.g. all-cause mortality or health-related quality of life). However, improvement in a surrogate outcome does not always correspond with an improvement in outcomes that are clinically meaningful and important to patients. Some examples of treatments that were implemented on positive effects on surrogate outcomes, but were later reversed when their effectiveness was examined with regard to clinically relevant outcomes, include: PCI for stable CAD, high-dose steroids for spinal cord injury, administration of calcium during cardiac arrest, cyclo-oxegenase-2 (COX-2) inhibitors for inflammation, a glycated haemoglobin (HbA1C) level of less than 7% for the management of diabetes, and bevacizumab for metastatic breast cancer survival.[22,36]

It is commonly assumed that the most expensive option is the best: that a higher price will result in better outcomes than a cheaper alternative, without considering the known value of the intervention.[37] This assumption can lead to unnecessary reversal because newer and more expensive practices are adopted to replace older and cheaper practices before the evidence has matured to support their use, leading to reversal when they are not found to be any better than the older or cheaper standard.[38]

There are also several characteristics of research that can lead to poor quality findings and an increased likelihood that the results are false or exaggerated, which in turn may lead to future reversal.[39] Increased financial interests or prejudices and the non-declaration of conflicts of interest are both established reasons for questioning the validity of research findings, as is novelty of a research field.[23,39] Novelty of a field in particular can create public pressure for early adoption of technologies that have not yet been fully tested. This public pressure often comes in the wake of sensationalized media

coverage of scientific breakthroughs which suggests that poor science reporting may also play a role in the premature adoption of practice, leading to higher risk for future reversal.[40,41]

The causes of reversal are not limited to suboptimal original research practices, as is the case in many of the examples above. Reversal should be a respected phenomenon and an expected element of scientific enquiry as new evidence emerges to contradict prior beliefs and standards. Medical and evidence reversal can occur because of newly discovered long-term side effects that could not have been known early in the course of a new treatment, even with well conducted trials: the kinds of side effects that require population-wide use over the longer term, as in Phase IV trials, to be discovered.[42] Reversals may occur over time because a practice that was a standard of care is no longer worth the cost because cheaper alternatives with similar effectiveness have since become available.[42] There is also a logistic issue surrounding the study of practices to the point of maturity before implementation, as the resources required to conduct multiple, large, clinical trials that follow patients for a sufficient length of time to determine the "true" effects of an intervention on patient-important outcomes make such a goal infeasible for some interventions. It is for these reasons that eliminating reversal entirely is impossible. However, reducing the impact of unnecessary reversals through preventing premature adoption of practices before a reasonable level of evidence has accrued would positively influence the harmful effects of reversal.

1.7 EXPLORING THE CHARACTERISTICS OF EVIDENCE REVERSAL

There are many factors that can contribute to the reversal of an established practice and many characteristics of research that can lead to the immature implementation of a practice before its true efficacy and effectiveness is understood. However, despite the knowledge that these causes of reversal exist, there are no frameworks that describe the characteristics of research that are associated with the contradiction of already established practices.[4,43,44]

'A decade of reversal: An analysis of 146 contradicted medical practices' by Prasad *et al.* was a major review of original research articles published in the New England Journal of Medicine (NEJM) between 2001 and 2010 that explored the prevalence of medical reversal in the medical literature.[4] They estimated the rate of contradiction over a 10-year period and found it to be approximately 40% of studies testing an existing practice.[4] They also provided descriptive statistics about the studies collected in their search – including the prevalence of various study designs, authors' conclusions, and the proportion of studies that tested medical practices that were new versus existing – and detailed qualitative descriptions of the 146 studies that they identified as reversals.[4]

While 'A decade of reversal' fulfilled a necessary and important step in moving towards a better understanding of this new field, the data provided was insufficient to accommodate an analytic assessment of the characteristics of research that may be associated with reversal. Prasad *et al.* conducted no quality assessments of the included studies, provided no details of study-level results (e.g. number of events and subjects in

study groups), and provided insufficient description of methodology with regards to their decision-making processes to facilitate reproducibility.[4]

However, these limitations provided a rationale for reproducing, updating, and expanding this review: using the results to create a database upon which to conduct a more quantitative analysis of the characteristics of reversal. The exploration of relevant characteristics in a database of trials with a logistic regression model – the outcome being the contradiction or reaffirmation of prior beliefs – will be one of the final steps in the development of a framework for identifying when established practices have been reversed: informing the framework with the strengths of associations that study characteristics may have with reversal.

In her thesis – the body of which laid much of the groundwork for this current thesis – Desirée Sutton proposed a framework of reversibility.[45] This framework included several summary measures of study quality covering question design and methodology, reporting (i.e. PICOTS, ROB, and modified GRADE), and several other measures including: modified optimal information size, fragility index, study abstract conclusions, and the lengths of time from a trial's start to its registration and from completion to publication.[45] These elements will be informed by our analyses and the framework will be developed and adapted accordingly.

1.8 CONCLUSIONS, IMPACT, AND THESIS OUTLINE

Despite its complexity, there is a paucity of research concerning the characteristics of reversal. Reversal imposes several dangers to the wellbeing of the population and minimizing the impact of unnecessary reversal would have tangible benefits. Overall health could be improved as potentially harmful practices may be

stopped earlier or prevented from implementation, trust in the medical and scientific communities would improve as fewer practices are redacted and undergo de-investment and de-adoption, and unnecessary resource expenditure would be reduced.

The meta-research community has only recently begun to explore reversals in a clinical context, hence the use of the term "medical reversal." In this thesis we propose the term "evidence reversal" as a more appropriate general term for the phenomenon of contradictory findings, as well as proposing a framework to identify when a reversal of evidence has occurred and several key areas of future research in the field of reversal. This thesis provides an in-depth exploration of evidence reversal and the process of developing a framework of reversibility. This exploration will contribute to the field of evidence reversals by bringing together multiple themes – both philosophical and practical – into a cohesive whole. This framework will promote consistent use of terminology related to reversal and serve to provide guidance for researchers in designing robust trials, potentially decreasing the risk of reversal in the future.

This first chapter has provided an overview of the concepts and theories that will be explored in the following chapters. The second chapter provides an updated systematic review on the concept of reversal and how it has been explored in the literature. The third chapter presents the methodology for our update of 'A decade of reversal' and how data extraction was conducted to create a database of trials and their characteristics. The fourth chapter provides an in-depth discussion of the process that we used to build our regression model and develop a framework of reversibility – from conceptualization and the analysis plan, to building the framework. The fifth chapter will present all results of our analysis of the characteristics of reversal including reproducing the descriptive

statistics originally published by Prasad *et al.*, as well as our expanded analyses (logistic regression) and framework. The sixth chapter contains the discussion of our findings, as well as the strengths and remaining limitations. The seventh chapter presents the impact and possible future applications of our proposed framework of reversibility, as well as introducing an initial toolbox for future reversal research and presenting the conclusions of the overall thesis, wherein we review our findings and what we have learned from our various reviews.

CHAPTER 2

Unlocking Evidence Reversal – An updated systematic review of Evidence Reversal terminology

Riaz Qureshi

With special thanks to Desirée Sutton and Dr. Janet Martin

Note: This chapter of this thesis was written in tandem with an article for publication. As such, there are many similarities in the sections of the published paper and this chapter. However, I was able to expand this chapter and go into more detail than the paper due to space restrictions within the publication.

Citation: Sutton D, **Qureshi R**, & Martin J. (2016). "Evidence Reversal – when new evidence contradicts current claims: A systematic overview review." Accepted for publication in the Journal of Clinical Epidemiology, July 2017.

Chapter Summary: This chapter is a systematic review of the literature to explore how the phenomenon of evidence reversal has been talked about in the past. Terms and definitions related to the phenomenon are compiled and organized into four major areas of research in the field of reversal.

CHAPTER 2

2.0 INTRODUCTION

In 2011, Prasad and Cifu coined the term "medical reversal" for the phenomenon of new evidence for an established practice that is methodologically stronger than previously conducted research, finding that the clinical practice is less effective or more harmful than was originally believed.[1] When such evidence arises for an established practice, it is "reversed" and steps should be taken to initiate its de-adoption, or removal, from practice. A medical reversal does not mean that the practice must be removed in its entirety; it is much more common that a reversal will provide a contraindication for a particular use in a particular population. Hormone replacement therapy (HRT) and percutaneous coronary intervention (PCI) still have clinical validity in an appropriate population, but it is now widely recognized that HRT does not reduce the risk of cardiovascular disease among post-menopausal women, and evidence shows that PCI for stable CAD does not reduce the risk of future adverse cardiac events, even though these practices were once though to provide net benefit.[9,14]

While 'medical reversal' is an appropriate term for the phenomenon that it represents, it is conceptually clinically oriented and implies the cessation of the reversed practice. The term may not always be appropriate as reversal occurs in non-medical fields, such as public health, and a reversal of evidence does not guarantee de-adoption: many practices continue after the evidence for them is contradicted.[26,46] We propose a new term, "evidence reversal" (ER), to describe the phenomenon in both medical and non-medical fields and, more appropriately, when new evidence that is stronger than preceding evidence contradicts the established evidence for a practice.

Reversal in a medical or clinical context is very common.[1] The percentage of studies that investigate an established clinical practice and subsequently lead to reversal may be as high as 40%.[4] Although the term for the phenomenon is still being diffused throughout the medical community, all physicians are familiar with the concept of reversal through the ever-changing guidelines for clinical practice.

The desirable progression of medicine is for newer and better interventions to replace older and less effective interventions.[2] This replacement of therapies is ideal because it implies that at any given time, patients are given a standard of care that is the best available treatment at the time.[2] However, despite reversal of evidence and subsequent de-adoption being expected, there are four implied harms to patients and the health care system as a whole. The first implication of reversal in a medical context is that the patients who were treated with the reversed practice were placed at a greater risk for, or actually experienced, unnecessary harm for little to no benefit.[2] The second is a risk that health care resources are being wasted because treatments that are unnecessary or of low value are being utilized before they can be reversed.[2,3] The third is an undermining of the trust in the medical system that is held by the public and by those who practice medicine.[2] And the fourth risk associated with reversal is the difficulty in de-adopting established practices from the medical community.[2,28]

In considering the challenge of de-adoption, beyond the medical community there exists the same difficulty in removing a paradigm from common belief, no matter what evidence may arise to contradict it. An example of the difficulties inherent in removing a consistently and clearly disproven theory from the view of the public is the persistence of the belief that vaccines cause autism.[47] In spite of evidence to contradict and deny the

claim, the belief remains strong enough that there have been lowered rates of childhood vaccination and an increase in the incidence of vaccine-preventable diseases in recent years.[47]

In order to curtail the harms of unnecessary reversal, by minimizing early conclusions that are based on insufficient evidence thereby preventing the premature uptake of new practices and theories, we must first have an understanding of the way that the phenomenon has been discussed and researched. In this systematic overview review, we explore how the concepts surrounding evidence reversal have been explored in the medical and non-medical literature. We create a compendium of terminology and definitions that relate to evidence reversal with the goal of bringing a degree of cohesion to this new and largely un-explored field.

2.1 METHODS

In 2014, Sutton and colleagues conducted a systematic literature review of the terminology surrounding evidence reversal as part of her thesis.[45] This systematic review was conducted with the purpose of finding how reversal had been discussed thus far in the scientific literature. As the term itself is new, and meta-research on the subject of the reversal of practices is sparse, we thought it best to update the search by Sutton *et al.* to include any new material from the two years since the search was last conducted. As this was an updated systematic review we aimed to use the same search methodology that was used in July of 2014. To this end, almost identical search strategies were applied to the same databases and sources that were searched.

A modification to the original search strategies for PUBMED, OVID MEDLINE, and EMBASE databases was devised under the supervision of a medical librarian. The

modification was necessary as a small error in the strategies for these databases led to two terms – "result" and "disinvest*" – being left out of the searches. After discussion, it was decided that the use of the term "result" would return too many unrelated citations, but "disinvest" would be added to the searches for these databases because it was highly relevant to reversal and returned a small and manageable number of citations.

2.1.1 Search strategy

A systematic review of the two-year period from January 1st, 2014 until July 6th, 2016 was conducted of eight scientific and gray literature databases including: PUBMED, Ovid MEDLINE, EMBASE, CINAHL, Web of Science, the Dissertations and Thesis Database, The Canadian Health Research Collaboration (CHRC), and GOOGLE Scholar™. In addition to this systematic searching, hand searching was conducted for the last two years for 27 blogs and websites and six journals.

Databases were searched using a combination of relevant subject headings and keywords including: evidence-based practice, patient care management, guidelines, clinical practice, practice guideline, physician's practice pattern, evidence-based, and disinvest*. In addition, searching by proximity was utilized using terms such as publication, evidence, practice, guideline*, medical, standard, unexpected, or surprising, paired with terms like revers*, change, contradict*, divest, or de-implement*. For the full search strategies for each database including number of returned citations, please see APPENDIX A.

2.1.2 Screening and inclusion criteria

All returned citations were imported into EndNote for screening, except for those collected from the CHRC, which were imported into, and screened in, Microsoft Excel. Screening was conducted at three levels: title, abstract, and full text. Due to time

constraints and the breadth of the search strategy, which was developed for high sensitivity and inclusion, screening was not conducted in duplicate. All new citations retrieved for the update of the search were screened by RQ, and the citations returned by DS were not duplicated. However, agreement was reached for final inclusion of articles and disagreements were settled by discussion.

To be included in this systematic overview review, all articles had to meet two criteria: they must have made some reference to the process of reversal or related concepts – according to our operational definition of ER – and they must have been a review article. Systematic reviews, meta-analyses, evidence syntheses, reviews, and collections of studies were included. All reviews pertaining to the phenomenon of new and stronger evidence contradicting current practice were included. These included direct and indirect references to evidence reversal including: medical reversals, changes in clinical practice guidelines or standards of care, and the disinvestment, de-implementation, or de-adoption of practices.

2.1.3 Quality assessment

Data extraction was performed independently with two authors (RQ and DS) verifying the other's work on a random subset of articles. Discrepancies were resolved through discussion.

Quality assessment of included reviews was done using the AMSTAR rating tool. The AMSTAR (A Measurement Tool to Assess Systematic Reviews) rating system is an instrument for assessing the methodological quality of systematic reviews.[48] It consists of 11 items and has been validated as a reliable quality assessment tool.[49] The greater an article's score out of 11, the greater the confidence in the findings of that review or group of reviews.[48,49]

Two out of the eleven items on the AMSTAR instrument ("Appropriate pooling of findings" and "Likelihood of publication bias") were not applicable to the articles included in our review because pooling of results across different Populations, Interventions, Comparators, Outcomes, Timing and Study Designs (PICOTS) is illogical, therefore it was determined that deviations from high methodological practice would not appreciably bias results. As a result of this modification, the maximum number of points that a review could achieve on AMSTAR was nine.

2.2 RESULTS

Systematic searches of the scientific and grey literature databases, and all hand searching of journals, blogs, and websites yielded 8117 unique citations. These citations were screened for exclusion criteria, resulting in 27 articles selected for inclusion after title, abstract, and full text screening. Five of these articles (from 2014) were already found from the search conducted two years ago. Duplication of these articles was expected and their inclusion validates the replication of previous search methods. Because five of these articles were already found, they were no longer counted as a part of the results in this review update. Therefore, 22 new articles were identified for inclusion in the systematic review. A further 8 articles were collected after screening the bibliographies, cited by, and related articles of the 22 identified via the database searches. Therefore, 30 new articles were added to the overall review through this update.

Combining these 30 new articles with the 57 retrieved from the 2014 search resulted in a total of 87 articles for inclusion in the final review. Please refer to Figure 2.1 PRISMA Flow-Chart for details of the article selection process used for this systematic review update (2014-July 2016).

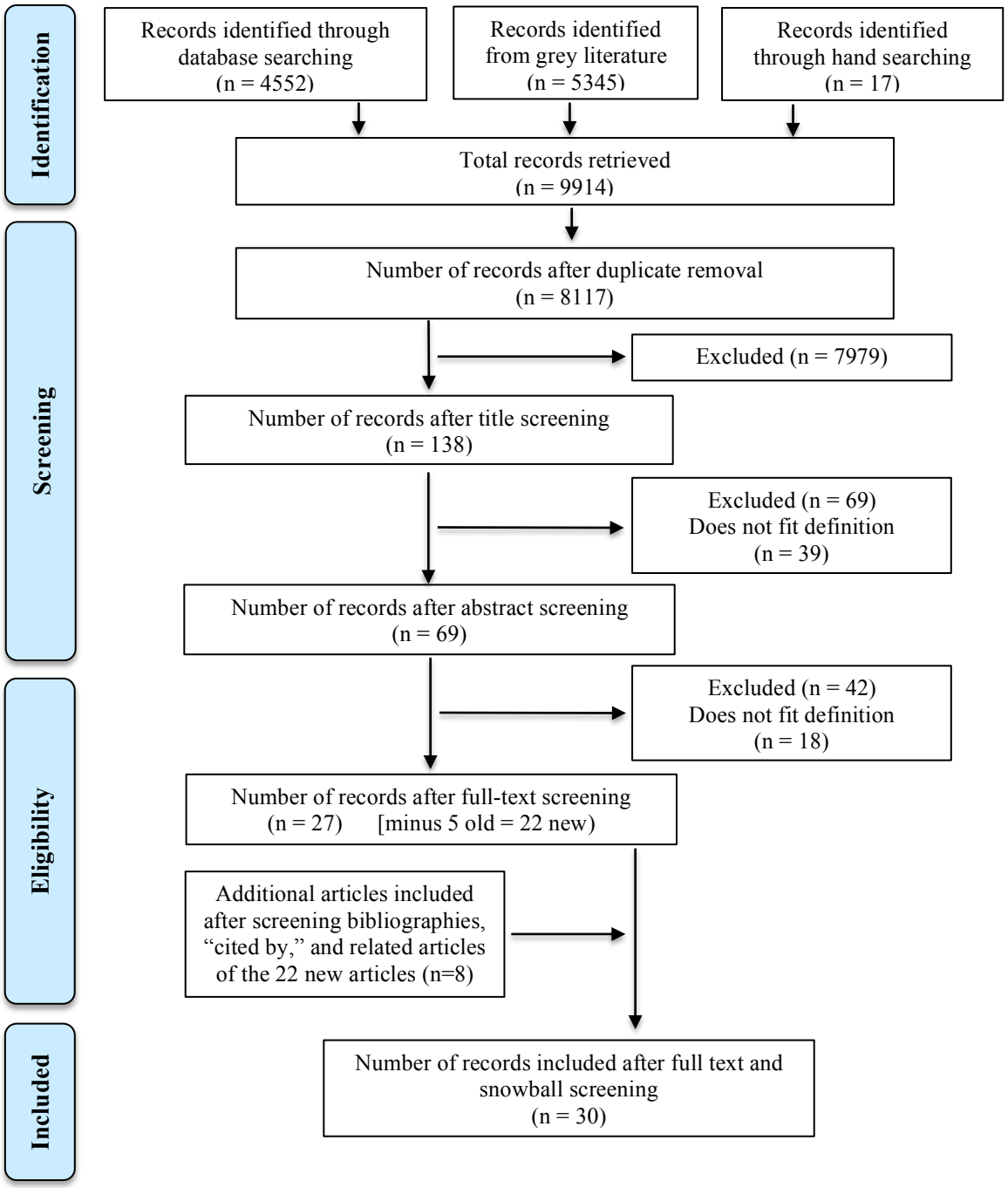


Figure 2.1 PRISMA flow diagram of study selection

The PRISMA flow-chart for the full review (i.e. the manuscript submitted for publication created by combining the original review and this update) may be found in Appendix A. Apart from the PRISMA diagram outlining the screening performed in this

update, the remaining results and discussion of findings presented are those of the full review, representing the findings from all 87 included articles.

2.2.1 Description of studies

The most common type of article that we found discussing evidence reversal was "collection of studies" (n = 58). These collections were mostly narrative and included letters to the editor, editorials, and recommendations for clinical practice based on new or important trial results. The majority of these collections of studies presented hand-picked or selected examples from journals or other reviews that the authors were discussing. Systematic reviews were the next most common type of article that we found (n = 24). The majority of articles have been published since 2011 (n = 71) with only a handful published in 2010 or before (n = 16).

Table 2.1 Characteristics of included studies

Characteristic	No. (%) of 87 Articles
Year of Publication	
2001-2005	7 (8)
2006-2010	9 (10)
2011-2015	59 (68)
2016	12 (14)
Type of Article	
Collection of Studies	58 (67)
Systematic Review	24 (28)
Overview Systematic Review	2 (2)
Systematic Scoping Review	1 (1)
Secondary Data Analysis and Review	1 (1)
Recursive Cumulative Meta-Analysis	1 (1)
Relationship to Reversal *	
Phenomenon of Reversal	32 (37)
Consequence of Reversal	35 (40)
Target of Reversal	79 (91)
Potential Predictor of Reversal	8 (9)
AMSTAR Quality Rating	
Very Low Quality (Score of 0 to 2)	63 (72)
Low Quality (Score of 3 to 5)	21 (24)
High Quality (Score of 6 to 8)	3 (3)
Very High Quality (Score of 9)	0 (0)

* Percentages do not sum to 100% due to the appearance of multiple terms within individual articles

2.2.2 Terminology related to evidence reversal

The operational definition for evidence reversal that was used to find relevant articles is: when new evidence – better powered, controlled, or designed than its predecessors – contradicts previously established claims. We found 50 unique sets of terms related to evidence reversal that we collated into four broader categories: a) terms for the phenomenon of ER, those that describe the event of new and better evidence contradicting older; b) terms for the consequences of reversal, the processes undertaken to remove a practice that has been reversed; c) terms for the targets of reversal or practices that are likely to be reversed in the future; and d) terms for potential predictors of ER. Table 2.2 presents the full list of identified terms and their definitions. Table 2.3 presents the frequency of the use of terms and how they relate to evidence reversal. APPENDIX B provides the complete data extraction for all included articles.

Table 2.2 Terms and associated definitions for evidence reversal

Year	Term used	Definition(s)
2001	Uncertainty [50]	"How much the treatment effect has changed over time and how much the pooled treatment effect will change the future." [50]
2002	No Articles	No Articles
2003	Discrepancy [51]	"Magnitude of the genetic effect as it changes over time." [51]
2004	Ineffective or harmful interventions [52]	"Treatments previously commonly practices, but not known to not work or cause harm." [52]
	Unfavourable or favourable shifts over time [53]	"Changes in whether results become less or more favourable for the experimental intervention over time." [53]
2005	Contradicted [8]	"Subsequent research contradicts efficacy claim." [8]
	Initially stronger effects [8]	"Subsequent research shows smaller magnitude of efficacy claim." [8]
	Proteus Phenomenon [54,55]	"Rapid, early succession of very contradictory conclusions." [54]
		"Extreme between-study opposing estimates of effect in the results of early studies followed by studies with diminishing between-study variance." [55]
2006	Contradicted [56]	"Diminishing effects for the strength of research findings and rapid alternations of exaggerated claims and extreme contradictions." [56]
	Proteus Phenomenon [56]	"Rapid alternation between exaggerated claims and extreme contradictions in early studies followed by studies with diminishing effects for the strength of research findings." [56]

2007	Change in evidence [57]	"Quantitative changes include differences of statistical significance or ≥50% effect change in magnitude for important outcomes. Qualitative changes include differences in definition of effectiveness, new data on harm, and caveats about previous evidence." [57]
	Disinvestment [58]	"The processes of withdrawing (partially or completely) resources from any existing healthcare practices, procedures, technologies or pharmaceuticals that are deemed to deliver little or no health gain relative to their cost, and thus are not efficient health resource allocations." [58]
	Inconsistency [59]	"Inconsistency occurs when there is large between-study heterogeneity (diversity) in the magnitude of the genetic effects." [59]
	Non-replication [59]	"Occurs when the GWA (Genome Wide Association) study proposes that there is a gene-disease association, but the accumulation of data from subsequent studies find no genetic effect." [59]
2008	Sacred Cows [31]	"Practices are considered routine and beyond dispute." [31]
		"A clinical practice despite research that shows that the practice is not helpful and may even be harmful to the patients we serve." [31]
2009	Assess new intervention – displace old [60]	"When a new intervention is presented to the relevant committee(s)† for regulatory assessment, and is considered a potential replacement for (an) established comparator(s) for that indication, then that comparator for that patient indication is automatically considered and assessed for disinvestment" [60]
	Class II Recommendation [61]	"Class II: conditions for which there is conflicting evidence and/or a divergence of opinion about the usefulness/efficacy of a procedure or treatment." [61]
		"Class IIa: weight of evidence/opinion is in favour of usefulness/efficacy." [61]
		"Class IIb: usefulness/efficacy is less well established by evidence/opinion." [61]
	Class III Recommendation [61]	"Class III: conditions for which there is evidence and/or general agreement that the procedure/treatment is not useful/effective and in some cases may be harmful." [61]
	False positive result [62]	"Report a treatment effect when in reality there is not effect." [62]
	Ineffective, harmful, or non-cost-effective interventions [60]	"Ineffective, non-cost-effective or harmful interventions." [60]
	Legacy items [60]	"Long-established technologies that have never had their cost-effectiveness assessed – look for coupling with other identification items. Automatically considered and assessed for disinvestment." [60]
	Technology development [60]	"When an intervention has evolved to the point that it differs markedly from the initial or prototype intervention that was originally assessed or funded, then the initial intervention should be reviewed (e.g., 256-slice compared with four-slice computed tomography)." [60]
2010	Snake oil [63]	"Nutritional supplements that are not worth it (inconclusive, slight, or no evidence of efficacy)." [63]
2011	False positive and inflated results [64]	"Report a treatment effect when in reality there is no effect." [64]
		"RRs were in opposite direction, larger, more than twice as large, more than 4 times as large, or different beyond chance in the highly cited vs. the largest study and in the highly cited study vs. the meta-analysis." [64]
	Medical reversal [1,2]	"The phenomenon of a new trial – superior to predecessors because of better design, increased power, or more appropriate controls – contradicting superior clinical practice." [1]

		"The phenomenon of a new trial – superior to predecessors because of better design, increased power, or more appropriate controls – contradicting superior clinical practice." [2]
		"A medical practice falls out of favour not by being surpassed, but when we discover that it did not work all along, either failing to achieve its intended goal or carrying harms that outweighed the benefits." [2]
	Recommendations for practice [65]	"Lack or presence of evidence" [65]
	Sacred cows [32]	"Practices are considered routine and beyond dispute." [32]
		"A clinical practice despite research that shows that the practice is not helpful and may even be harmful to the patients we serve." [32]
2012	Disinvestment [28]	"The process of withdrawing health resources from any existing healthcare practices, procedures, technologies and pharmaceuticals that are deemed to deliver no or low health gain for their cost and thus [do] not [represent] efficient health resource allocation." [28]
		"The cessation or restriction of potentially harmful, clinically ineffective or cost inefficient practices." [28]
		"The process of taking resources from one service in order to use them for other purposes (i.e. reallocation of resources)." [28]
	Improper use [28]	"Any existing healthcare practices, procedures, technologies and pharmaceuticals that are deemed to deliver no or low health gain for their cost and thus [do] not [represent] efficient health resource allocation'." [28]
	Low-value practices [66]	"Ineffective and/or unsafe services, treatments not proven to be clinically effective." [66]
	Medical reversal [3,26]	"Established standards must be abandoned not because a better replacement has been identified but simply because what was thought to be beneficial was not." [3]
		"Oftentimes, years after a practice was introduced, the medical community puts it to the test in large, well done randomized trials. Empirical evidence suggests that when this happens, nearly half of those practices are contradicted. We call this phenomenon 'medical reversal'." [26]
	Negative list [28]	"[practices] that have been superseded or demonstrated to be ineffective or harmful." [28]
	Obsolete/outmoded/abandoned technologies [28]	"Those that have been superseded or demonstrated to be ineffective or harmful." [28]
	POEM likely to change clinical practice [67]	"A study that is valid (avoids important biases), reports patient-important outcomes (such as morbidity, mortality, or quality of life) and changes clinical practice." [67]
	Research updates most likely to change clinical practice [68,69]	"Research updates most likely to change clinical practice." [68,69]
	Services not medically necessary [28]	"Any existing healthcare practices, procedures, technologies and pharmaceuticals that are deemed to deliver no or low health gain for their cost and thus [do] not [represent] efficient health resource allocation." [28]
	Things providers and patients should question [70]	"Wasteful or unnecessary medical tests, treatments and procedures" [70]
	Unnecessary medical tests, treatments, and procedures [70]	"Wasteful or unnecessary medical tests, treatments and procedures" [70]

	Wasteful medical tests, treatments, and procedures [70]	"Wasteful or unnecessary medical tests, treatments and procedures" [70]
2013	Disinvestment [71]	"The complete or partial withdrawal of resources from healthcare services and technologies that are regarded as unsafe, ineffective or inefficient." [71]
	Ineffective technologies [71]	"Healthcare services and technologies that are regarded as unsafe, ineffective or inefficient." [71]
	Low-value practices / health care [72–76]	"Clinical decisions that are of little value to patients, amenable to improvement through standardization, and actionable by front-line providers." [72]
		"Ineffective or lack evidence on their effectiveness, negative risk-benefit balance, more cost-effective alternatives exists, obsolete due to the introduction of new technologies." [73]
		"Interventions that robust evidence reveals are of no benefit, or even harmful." [74]
		"Health care services that provide little or no benefit – whether through overuse or misuse." [75]
		"…Not clinically effective for a given indication. It may be unsafe for everyone, or for subgroups of patients with risk factors. It may have a poor risk-benefit profile overall, or when used inappropriately." [76]
	Medical reversal [4,35,36,77]	"Reversal was designated when a current medical practice was found to be inferior to a lesser or prior standard." [4]
		"The phenomenon of a new superior trial that contradicts current clinical practice." [36]
		"Medical reversal happens when new trials – better powered, designed or controlled than predecessors – contradict current standard of care." [77]
		"Modifications or even retractions, of important medical practice recommendation… [which] challenge traditional medical opinion." [35]
	Obsolete technologies [71]	"Healthcare services and technologies that are regarded as unsafe, ineffective or inefficient." [71]
	Overdiagnosis [78]	"Waste of resources on unnecessary care" [78]
	Overused or misused tests and treatments [79]	"Unnecessary tests and procedures that don't benefit the patient and can even cause harm." [79]
	POEM likely to change clinical practice [80]	"A study that is valid (avoids important biases), reports patient-important outcomes (such as morbidity, mortality, or quality of life) and changes clinical practice." [80]
	Research updates most likely to change practice [81–84]	"Research updates most likely to change clinical practice." [81–84]
	Sacred cows [33]	"Practices are considered routine and beyond dispute." [33]
		"A clinical practice despite research that shows that the practice is not helpful and may even be harmful to the patients we serve." [33]
	Snake oil [85]	"Nutritional supplements that are not worth it (inconclusive, slight or no evidence of efficacy)." [85]
	Too much medicine [78]	"Waste of resources on unnecessary care" [78]
	Unproven therapies [34]	"No proven value by current Grading of Recommendations Assessment, Development and Evaluation (GRADE) guidelines, US Preventive Task Force Services criteria, or other similar criteria." [34]
	Waste [75]	" Inappropriate overuse of an otherwise effective intervention." [75]

2014	Change in treatment guidelines [86]	"New RCT findings." [86]
	Contradicted established medical practices [44]	"When large, well-done randomized trials have contradicted current medical practice." [44]
	De-implementation [44]	"Abandonment of medical interventions." [44]
		"Stopping practices that are not evidence-based." [44]
	Inappropriate care [87]	"That relating to the use or non-use of a health service intervention based on the evaluation of (a) evidence of effectiveness; and/or (b) economic implications; and/or (c) other health system impacts; and/or (d) consideration of ethical implications and societal values." [87]
	Medical reversal [88]	"A phenomenon in which 'a medical practice is found to be inferior to some lesser or prior standard of care.'" [88]
	POEM likely to change clinical practice [89]	"A study that is valid (avoids important biases), reports patient-important outcomes (such as morbidity, mortality, or quality of life) and changes clinical practice." [89]
	[Sacred Cows] Practices not supported by the evidence [30]	"Practices are considered routine and beyond dispute." [30]
		"A clinical practice despite research that shows that the practice is not helpful and may even be harmful to the patients we serve." [30]
	Snake oil [90]	"Nutritional supplements that are not worth it (inconclusive, slight or no evidence of efficacy)." [90]
	Research updates most likely to change practice [91–93]	"Research updates most likely to change clinical practice." [91–93]
	Research waste [94]	"Avoidable waste or inefficiency in biomedical research." [94]
	Unnecessary treatments, tests, and procedures [95]	"Do not add value to care … potentially expose patients to harm, leading to more testing to investigate false positives and contributing to stress for patients … increased strain on the resources of our health care system." [95]
	Unproven medical practice [30]	"Many medical practices are largely untested or have insufficient evidence unable to support or refute interventions." [30]
2015	Abandonment / abandon* [96,97]	"Processes of withdrawing (partially or completely) health resources from any existing health care practices, procedures, technologies or pharmaceuticals that are deemed to deliver little or no health gain for their cost, and are thus not efficient health resource allocations." [96]
		"The discontinuation of a clinical practice after it was previously adopted." [97]
	[Change / decline / change / drop in] in use [96,97]	"Processes of withdrawing (partially or completely) health resources from any existing health care practices, procedures, technologies or pharmaceuticals that are deemed to deliver little or no health gain for their cost, and are thus not efficient health resource allocations." [96]
		"The discontinuation of a clinical practice after it was previously adopted." [97]
	Contradict* [96,97]	"Processes of withdrawing (partially or completely) health resources from any existing health care practices, procedures, technologies or pharmaceuticals that are deemed to deliver little or no health gain for their cost, and are thus not efficient health resource allocations." [96]
		"The discontinuation of a clinical practice after it was previously adopted." [97]
	Clinical redesign or re-prioritization [97]	"The discontinuation of a clinical practice after it was previously adopted." [97]

	De-adoption / de-adopt* [96,97]	"Processes of withdrawing (partially or completely) health resources from any existing health care practices, procedures, technologies or pharmaceuticals that are deemed to deliver little or no health gain for their cost, and are thus not efficient health resource allocations." [96]
		"The discontinuation of a clinical practice after it was previously adopted." [97]
	De-commission / de-list [97]	"The discontinuation of a clinical practice after it was previously adopted." [97]
	Decrease use / reduc* [96,97]	"Processes of withdrawing (partially or completely) health resources from any existing health care practices, procedures, technologies or pharmaceuticals that are deemed to deliver little or no health gain for their cost, and are thus not efficient health resource allocations." [96]
		"The discontinuation of a clinical practice after it was previously adopted." [97]
	De-funding or resource release [97]	"The discontinuation of a clinical practice after it was previously adopted." [97]
	Discontinuation / discontinu* [96,97]	"Processes of withdrawing (partially or completely) health resources from any existing health care practices, procedures, technologies or pharmaceuticals that are deemed to deliver little or no health gain for their cost, and are thus not efficient health resource allocations." [96]
		"The discontinuation of a clinical practice after it was previously adopted." [97]
	De-implementation / de-implement* [97,98]	"Can involve overuse, underuse, and/or misuse of health services, products, and resources." [98]
		"The discontinuation of a clinical practice after it was previously adopted." [97]
	Disinvestment / disinvest [96,97,99]	"Processes of withdrawing (partially or completely) health resources from any existing health care practices, procedures, technologies or pharmaceuticals that are deemed to deliver little or no health gain for their cost, and are thus not efficient health resource allocations." [96]
		"The process of (partially or completely) withdrawing health resources from any existing healthcare practices, procedures, technologies, or pharmaceuticals that are deemed to deliver little or no health gain for their cost, and thus are not efficient health resource allocations." [99]
		"The discontinuation of a clinical practice after it was previously adopted." [97]
	Do-not-do [97]	"The discontinuation of a clinical practice after it was previously adopted." [97]
	Harmful practices [98,100]	"Can involve overuse, underuse, and/or misuse of health services, products, and resources." [98]
		"Low value, unnecessary, or harmful to patients." [100]
	Inappropriate care [101]	"Treatments that evidence clearly shows should not be done routinely, or at all." [101]
	Ineffective [technology / practice] [97,99]	"The discontinuation of a clinical practice after it was previously adopted." [97]
		"Ineffective technologies are usually identified by evaluating their effectiveness, safety, and cost-effectiveness. In addition, overuse or misuse of technologies can lead to ineffectiveness." [99]
	Low-value health care / services / practices [97,98,100,102]	"The discontinuation of a clinical practice after it was previously adopted." [97]
		"Care that is avoidable/not necessary/of low value." [102]
		"Can involve overuse, underuse, and/or misuse of health services, products, and resources." [98]
		"Low value, unnecessary, or harmful to patients." [100]

	Medical reversal [96,97,103–105]	"Processes of withdrawing (partially or completely) health resources from any existing health care practices, procedures, technologies or pharmaceuticals that are deemed to deliver little or no health gain for their cost, and are thus not efficient health resource allocations." [96]
		"When an accepted practice – a diagnostic test, medication, or procedure – is overturned. The practice is not replaced by something better, but shown to be inferior to a pre-existing, less intensive, or less invasive one." [103]
		"Reversal of medical practice requiring significant changes in standards of care, workflow, and decision making." [104]
		"The discontinuation of a clinical practice after it was previously adopted." [97]
		"When a current practice is found to be no better than, or inferior to, a prior standard." [105]
	Misuse [96,97]	"Processes of withdrawing (partially or completely) health resources from any existing health care practices, procedures, technologies or pharmaceuticals that are deemed to deliver little or no health gain for their cost, and are thus not efficient health resource allocations." [96]
		"The discontinuation of a clinical practice after it was previously adopted." [97]
	Obsole* [96,97,99]	"Processes of withdrawing (partially or completely) health resources from any existing health care practices, procedures, technologies or pharmaceuticals that are deemed to deliver little or no health gain for their cost, and are thus not efficient health resource allocations." [96]
		"The discontinuation of a clinical practice after it was previously adopted." [97]
		"Obsolete technology: Any health technology in use for one or more indications, whose clinical benefit, safety, and/or cost-effectiveness have been significantly superseded by other available alternatives or are not supported by evidence." [99]
	Opportunity cost [96]	"Processes of withdrawing (partially or completely) health resources from any existing health care practices, procedures, technologies or pharmaceuticals that are deemed to deliver little or no health gain for their cost, and are thus not efficient health resource allocations." [96]
	Overdiagnosis [96]	"Processes of withdrawing (partially or completely) health resources from any existing health care practices, procedures, technologies or pharmaceuticals that are deemed to deliver little or no health gain for their cost, and are thus not efficient health resource allocations." [96]
	Overtreatment / overmedicalization / Medical over use [96,97,106,107]	"Processes of withdrawing (partially or completely) health resources from any existing health care practices, procedures, technologies or pharmaceuticals that are deemed to deliver little or no health gain for their cost, and are thus not efficient health resource allocations." [96]
		"The provision of health care when the "risk of harm exceeds its potential benefit," when the benefits are negligible, or when fully informed patients would forego care." [107]
		"Treatment of overdiagnosed conditions, or treatment that has minimal evidence of benefit or is excessive (in complexity, duration, or cost) relative to alternative accepted standards." [106,107]
		"The discontinuation of a clinical practice after it was previously adopted." [97]
	POEM likely to change clinical practice [108]	"A study that is valid (avoids important biases), reports patient-important outcomes (such as morbidity, mortality, or quality of life) and changes clinical practice." [108]

	Practice-changing evidence [109]	"Potential for practice change." [109]
	Reappraisal [96,97]	"Processes of withdrawing (partially or completely) health resources from any existing health care practices, procedures, technologies or pharmaceuticals that are deemed to deliver little or no health gain for their cost, and are thus not efficient health resource allocations." [96]
		"The discontinuation of a clinical practice after it was previously adopted." [97]
	Reassess* or [evidence-based / health technology] reassessment [96,97]	"Processes of withdrawing (partially or completely) health resources from any existing health care practices, procedures, technologies or pharmaceuticals that are deemed to deliver little or no health gain for their cost, and are thus not efficient health resource allocations." [96]
		"The discontinuation of a clinical practice after it was previously adopted." [97]
	Recommendations for practice [110]	"New RCT findings." [110]
	Redeploy [97]	"The discontinuation of a clinical practice after it was previously adopted." [97]
	Refute [97]	"The discontinuation of a clinical practice after it was previously adopted." [97]
	Re-invest or substitutional re-investment [97]	"The discontinuation of a clinical practice after it was previously adopted." [97]
	Relinquish* [97]	"The discontinuation of a clinical practice after it was previously adopted." [97]
	Remov* [97]	"The discontinuation of a clinical practice after it was previously adopted." [97]
	Replace [97]	"The discontinuation of a clinical practice after it was previously adopted." [97]
	Resource re-allocation [96,97]	"Processes of withdrawing (partially or completely) health resources from any existing health care practices, procedures, technologies or pharmaceuticals that are deemed to deliver little or no health gain for their cost, and are thus not efficient health resource allocations." [96]
		"The discontinuation of a clinical practice after it was previously adopted." [97]
	Sacred Cows [111]	"Clinical practices so rooted in tradition that they are resistant to change, despite reduced quality of care, patient outcomes, and greater costs compared with newer practices." [111]
	Selective disinvestment [98]	"Can involve overuse, underuse, and/or misuse of health services, products, and resources." [98]
	Stop* [97]	"The discontinuation of a clinical practice after it was previously adopted." [97]
	Undiffusion [112]	"Abandoning established practices found to be ineffective, disruptive, or the cause of net harm — or when better practices come along." [112]
	Waste [96]	"Processes of withdrawing (partially or completely) health resources from any existing health care practices, procedures, technologies or pharmaceuticals that are deemed to deliver little or no health gain for their cost, and are thus not efficient health resource allocations." [96]
	Withdraw* or withdrawing from a service and redeploying resources [97]	"The discontinuation of a clinical practice after it was previously adopted." [97]

2016 *As of July 24th, 2016	Do-not-Do recommendations [113]	"Clinical practices, identified during the development of guidance that should be discontinued or not used routinely." [113]
	Grade D Recommendation [114]	"The USPSTF recommends against the service. There is moderate or high certainty that the service has no net benefit or that the harms outweigh the benefits. Discourage the use of this service." [114]
	Ineffective services [115]	"Possibly ineffective or low-value services." [115]
	I statement [114]	"The USPSTF concludes that the current evidence is insufficient to assess the balance of benefits and harms of the service. Evidence is lacking, of poor quality, or conflicting, and the balance of benefits and harms cannot be determined" [114]
	Known uncertainty [116]	"Uncertainties about the effects of treatments." [116]
	Low-value health care [115]	"Possibly ineffective or low-value services." [115]
	Medical reversal [117,118]	"An accepted medical practice, often widely adopted, that is later found to be no better or worse than a previous standard of care." [117]
		"Medical reversals occur when the results of preclinical, observational and/or early phase studies fail to predict the results of subsequent randomized clinical trials, but the practice has already gained widespread acceptance." [118]
	POEM likely to change clinical practice [119]	"A study that is valid (avoids important biases), reports patient-important outcomes (such as morbidity, mortality, or quality of life) and changes clinical practice." [119]
	Practice-changing evidence [120]	"Potential for practice change." [120]
	Translation failure [118]	"Translation failure occurs when the results of preclinical, observational and/or early phase studies fail to predict the results of well done (i.e. appropriately controlled, adequately powered, and properly conducted) phase III or randomised clinical trials." [118]
	Trials likely to change practice [121]	"New RCT findings." [121]
	Unnecessary waste [115]	"Ineffective or low- value services, which are possibly provided at the expense of the social health insurance in Austrian primary care." [115]

Sixteen reviews used the term "Medical Reversal", and the definitions that were presented were nearly identical to our operational definition for evidence reversal.[1–4,26,35,36,77,88,96,97,103–105,117,118] The use of the term "contradicted," in the context of established medical practices, also had a similar definition, but with a focus on the comparison of initial conceptual studies and subsequent research, as opposed to evidence reversal which compares new evidence to previous evidence.[8,44,97] Another term for the comparison of initial conceptual, and often extreme, results with subsequent research – but one that is used exclusively in the context of genome wide association studies – is the

"Proteus Phenomenon."[54–56] Other terms that we found referring to the phenomenon of ER include: "change in evidence" [57]; "opportunity cost" [96]; "overdiagnosis" [78,96]; "practice changing evidence" [109,120]; "refute" [97]; "translation failure" [118]; and "favourable / unfavourable shifts over time."[53] Each of these terms describes the phenomenon of new evidence coming to light that changes a previously held belief about an established practice.

Separate from terms that describe the phenomenon of evidence reversal are terms for the processes of reversal, or the actions that should take place after a reversal has occurred. These terms include: "abandonment" [96,97]; "assess new interventions – displace old or replace" [60,97]; "change in treatment [guidelines / practice]" [86,97,104]; "clinical redesign or reprioritization" [97]; "de-adoption or dis-adoption" [96,97]; "decommission or de-list" [97]; "[decrease / decline / change / drop in] use" [96,97]; "defunding or resource release" [97]; "de-implementation" [44,97,98]; "discontinuation" [96,97]; "disinvestment" [28,58,71,96–99]; "overtreatment, medical overuse, overuse, misuse, or 'too much medicine'" [78,96,97,106,107]; "reassess*, [evidence-based / health technology] reassessment, or re-appraisal" [96,97]; "recommendations for practice" [65,110]; "re-invest, substitutional re-investment, re-allocation, or redeploy" [96,97]; "remov* or stop*"[97]; "undiffusion" [112]; and "withdraw, withdrawing from a service and redeploying resources, or relinquish."[97]

Another category of terms that are related to reversal, but do not refer to the phenomenon or its consequences are those that describe targets. These terms refer to practices that are known to be reversals or are likely to be reversed and are therefore targeted for removal. These terms include: "Class II / IIa / IIb/ III recommendation" [61]; "do-not-do recommendations" [97,113]; "grade D recommendations" [114]; "I statement" [114];

"[inappropriate / improper] [care / use]" [28,87,97,98,101]; "[Ineffective / harmful / non-cost-effective] [interventions / technologies]" [52,60,71,97–100,115]; "legacy items" [60]; "low-value [practices / health-care / services / intervention]" [66,72–76,97,98,100,102,115]; "negative list" [28]; "[obsolete / outmoded / abandoned] technologies" [28,71,96,97,99]; "[overused / misused] tests and treatments" [79]; "[research updates most / trials / POEMS] likely to change clinical practice" [67–69,80–84,89,91–93,108,119,121]; "sacred cows" [30–33,111]; "snake oil" [63,85,90]; "technology development" [60]; "things providers and patients should question or practices not supported by evidence" [30,70]; "unnecessary medical [tests / treatments / procedures], unnecessary [tests / treatments / procedures], or services not medically necessary" [28,70,95]; "unproven [therapies / medical practice]" [34,44]; and "waste, research waste, unnecessary waste, or wasteful medical tests, treatments, and procedures."[70,75,94,96,115]

The final category of terms that we propose is related to evidence reversal is potential predictors of future reversal. These terms all refer to red flags in clinical research: their presence could bring the efficacy and strength of the evidence surrounding the investigated practice into question. Potential predictors of future reversal include: "[Discrepancy / inconsistency / uncertainty] or known uncertainty" [50,51,59,116]; "[false positive / inflated] results" [62,64]; "initially stronger effects" [8]; and "non-replication."[59]

Table 2.3 Frequency of terms and their relation to evidence reversal

Term Set [a]	No. (%) of 87 Articles [b]	Year of First Appearance	Relationship to Reversal	References
Abandonment / abandon*	2 (2)	2015	Consequence	[96,97]
Assess new interventions – displace old or replace	2 (2)	2009	Consequence	[60,97]
Change in evidence	1 (1)	2007	Phenomenon	[57]
Change in treatment guidelines / practice	3 (3)	2014	Consequence	[86,97,104]
Class II recommendation	1 (1)	2009	Target	[61]
Class III recommendation	1 (1)	2009	Target	[61]
Clinical redesign or re-prioritization	1 (1)	2015	Consequence	[97]

Contradicted / contradict* / refute / contradictory result	5 (6)	2005	Phenomenon	[8,44,56,96,97]
De-adoption / de-adopt* / dis-adoption	2 (2)	2015	Consequence	[96,97]
Decommission / de-list	1 (1)	2015	Consequence	[97]
[Decrease / decline / change / drop in] use or reduc*	2 (2)	2015	Consequence	[96,97]
Defunding or resource release	1 (1)	2015	Consequence	[97]
De-implementation / de-implement*	3 (3)	2014	Consequence	[44,97,98]
Discontinuation / discontinu*	2 (2)	2015	Consequence	[96,97]
Discrepancy / inconsistency / uncertainty or known uncertainty	4 (5)	2003	Potential Predictor	[50,51,59,116]
Disinvestment / disinvest*	7 (8)	2007	Consequence	[28,58,71,96–99]
Do-not-do recommendations	2 (2)	2007	Target	[97,113]
[False positive / inflated] results	2 (2)	2009	Potential Predictor	[62,64]
Grade D recommendations	1 (1)	2016	Target	[114]
I statement	1 (1)	2016	Target	[114]
[Inappropriate / improper] [care / use]	5 (6)	2014	Target	[28,87,97,98,101]
[Ineffective / harmful / non-cost-effective] [interventions / technologies / practices]	8 (9)	2004	Target	[52,60,71,97–100,115]
Initially stronger effects	1 (1)	2005	Potential Predictor	[8]
Legacy items	1 (1)	2009	Target	[60]
Low-value [practices / health care / services / intervention]	11 (13)	2012	Target	[66,72–76,97,98,100,102,115]
Medical reversal / reversal	16 (18)	2011	Phenomenon	[1–4,26,35,36,77,88,96,97,103–105,117,118]
Negative list	1 (1)	2012	Target	[28]
Non-replication	1 (1)	2007	Potential Predictor	[59]
[Obsolete / outmoded / abandoned] technologies	5 (6)	2012	Target	[28,71,96,97,99]
Opportunity cost	1 (1)	2015	Phenomenon	[96]
Overdiagnosis	2 (2)	2013	Phenomenon	[78,96]
Overtreatment or medical overuse or “too much medicine” or overuse or misuse	5 (6)	2013	Consequence	[78,96,97,106,107]
[Overused / misused] tests and treatments	1 (1)	2013	Target	[79]
Practice changing evidence	2 (2)	2015	Phenomenon	[109,120]
Proteus phenomenon	3 (3)	2005	Phenomenon	[54–56]
Reassess* or [evidence-based / health technology] reassessment or re-appraisal	2 (2)	2015	Consequence	[96,97]

Recommendations for practice	2 (2)	2015	Consequence	[65,110]
Re-invest or re-allocation or substitutional re-investment or redeploy	2 (2)	2015	Consequence	[96,97]
Remov* or Stop*	1 (1)	2015	Consequence	[97]
[Research updates most / trials / POEMS] likely to change clinical practice	15 (17)	2012	Target	[67–69,80–84,89,91–93,108,119,121]
Sacred cows	5 (6)	2008	Target	[30–33,111]
Snake oil	3 (3)	2010	Target	[63,85,90]
Technology development	1 (1)	2009	Target	[60]
Things providers and patients should question or practices not supported by evidence	2 (2)	2012	Target	[30,70]
Translation failure	1 (1)	2016	Phenomenon	[118]
Undiffusion	1 (1)	2015	Consequence	[112]
Unfavourable or favourable shifts over time	1 (1)	2004	Phenomenon	[53]
Unnecessary medical [tests / treatments / procedures] or unnecessary [tests / treatments / procedures] or services not medically necessary	3 (3)	2012	Target	[28,70,95]
Unproven [therapies / medical practice]	2 (2)	2013	Target	[34,44]
Waste or research waste or unnecessary waste or wasteful medical tests, treatments and procedures	4 (5)	2012	Target	[70,75,94,96,115]
Withdraw* or withdrawing from a service and redeploying resources or relinquish	1 (1)	2015	Consequence	[97]

[a] *wildcard notation signifies multiple endings for the given term
[b] Percentages do not total 100 due to the appearance of multiple terms within individual articles

2.2.3 Quality Assessment

The overall confidence in findings of included articles was very low with a mode score of '1', a median score of '2', and a mean score of '2'. Most of the included articles declared conflicts of interest, but none of them described the potential conflicts of the examples that they presented. This led to the "conflicts of interest" item being uniformly not present among the included articles. The next four AMSTAR items that were the least present among included articles were: "list of included and excluded studies" (1%), "publication status in inclusion" (8%), "quality of included studies" (11%), and

"appropriate conclusions" (11%). The most present AMSTAR item among included articles was "characteristics of included studies" (85%). The majority of articles presented study characteristics in a non-table format (n = 47). The two AMSTAR items that had the greatest uncertainty were "study selection and data extraction in duplicate" and "comprehensive literature search," the presence of which was unclear in 60% of included articles. APPENDIX C contains the full AMSTAR evaluation for all included articles.

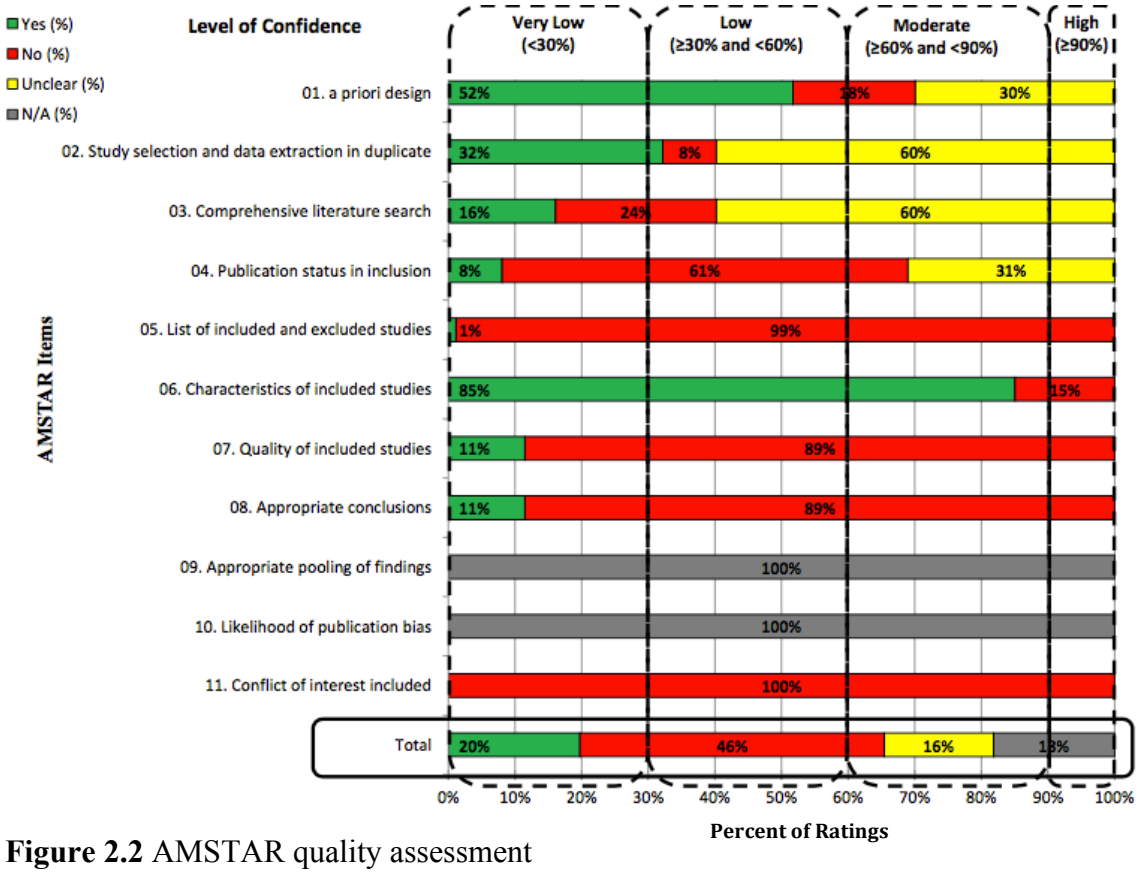


Figure 2.2 AMSTAR quality assessment

2.3 DISCUSSION

In this systematic overview review, we expected to find a wide range of terms for the phenomenon of evidence reversal because the term "Medical Reversal" has only been

in use since 2011, but the phenomenon has been present for longer.[1] We found 50 unique sets of terms. The true number of terms was greater (179), but some terms were synonymous and were therefore grouped together (e.g. "[ineffective / harmful / non-cost-effective] [interventions / technologies]" are six possible unique terms, but one unique set). Furthermore, we were interested to find that not all of these sets of terms had the same relation to evidence reversal. We set out to understand how the phenomenon had been explored thus far, both inside and outside of the academic literature, and based on our findings we propose that there are four essential facets to the evidence reversal discussion and meta-research.

The first facet is research about the phenomenon itself: research surrounding the event of new evidence contradicting what was previously found and believed about a given practice or theory. The second facet is research around the processes and consequences of evidence reversal: the difficulties inherent in, and methods by which practices are de-implemented. The third facet of studying evidence reversal is identifying and declaring the targets for evidence reversal: the practices, health-care, and services that are known or suspected to be of low value. The fourth facet of this area of meta-research is studying the potential predictors of future reversal: the characteristics of research that may lead to reversal of a practice in the future.

In order to reduce the impact of evidence reversals, the two most important areas of research are the latter two facets: identifying the targets of ER for removal and the potential predictors of ER for prevention. These areas of meta-research will take time and a concerted effort on behalf of the meta-research community. The first step to reduce the

impact is to bring cohesion to the first two facets of ER: studying the phenomenon of ER as well as the processes and consequences of ER.

The 26 unique sets of terms describing the phenomenon and consequences of ER suggest that there is currently no cohesion in this field of meta-research. While terms such as "Medical Reversal" and "De-implementation" are used most often, there are many other terms that have similar definitions but have slight contextual differences. For example: "Unfavourable or favourable shifts over time," defined as "Changes in whether results become less or more favourable for the experimental intervention over time" [53]; "Change in Evidence," defined as "Quantitative changes include differences of statistical significance or ≥50% effect change in magnitude for important outcomes. Qualitative changes include differences in definition of effectiveness, new data on harm, and caveats about previous evidence" [57]; or "Translation Failure," defined as "When the results of preclinical, observational and/or early phase studies fail to predict the results of well-done (i.e. appropriately controlled, adequately powered, and properly conducted) phase III or randomised clinical trials."[118] This variety of terms causes unnecessary confusion and increases the difficulty for those who wish to understand the processes involved. This difficulty was evident while designing database specific search strategies as there are no subject headings that are specific to evidence reversal.

In 2015, Niven *et al.* proposed the term "De-adoption" as a unifying term for the process of removing a practice from use. We agree that this would be a good term to describe the consequences of reversal in an all-encompassing context, as opposed to the most frequently cited term for the process – disinvestment – which was more often used in a monetary/economical context. In adopting the term "de-adoption" for the process and

consequences of evidence reversal, there is still the confusion created by the many terms for the phenomenon. Although it has not been used before in the literature, we propose the term *Evidence Reversal* as a unifying term because it represents finding contradictory evidence for any established practices or beliefs in both medical and non-medical fields, unlike the current most commonly used term, "Medical Reversal," which is restricted to the field of medicine and clinical practice.

2.3.1 Significance and future directions

The significance of this review for the field and for this thesis lies in the proposed unity that the term Evidence Reversal would bring to this area of meta-research. Previous research in the field of reversals has been focused on the medical literature and clinical practices. As a term for the phenomenon, "medical reversal" is clinically oriented and carries an implication of cessation of practice. This systematic overview review expands the definition from "medical reversal" to "evidence reversal," thereby encompassing both medical and non-medical practices, and providing an appropriate term for when the evidence has been reversed, but the practice continues to be used.

The next step for resolving the large collection of terms and definitions would be to form a common language framework using input from content experts in medical reversal, disinvestment, and meta-research through the Delphi Method to reach consensus on subject headings and how they should be defined. Once created, the framework will promote consistent use of terms and concepts to maximize comparability, repeatability, and quality of evidence, which will allow universal discussion and higher quality reviews in the future, thus advancing the field of evidence reversal in a more structured way.

This review will inform the development of such a framework because it has revealed that although exploration of reversals as a field of research has only recently

begun, there are many different terms that have similar definitions. This discrepancy suggests that the best way to approach ER may be through a simplified definition such as: "when newer and stronger evidence contradicts a previously held belief that was based on older or weaker evidence."

2.3.2 Strengths and limitations

This review was very thorough and the search strategies and methodologies employed gave the results a very high sensitivity to capture all of the relevant material. As the purpose of this review was to explore the literature and capture the different ways that reversal has been described, our broadly defined inclusion criteria allowed us to capture what should be the vast majority of relevant articles related to evidence reversal.

The high sensitivity of our search is also one of its weaknesses. This review was very labour intensive and thus, screening was not conducted in duplicate and inclusion may have been more subjective than is typically desirable for a systematic review. While there was a six-month period of overlap between the original searches and this update (from January 2014 until July 2014), agreement was not calculated for any of the screening levels (i.e. title, abstract, full-text). However, the articles for final inclusion were agreed upon between authors (RQ and DS), with all disagreements resolved through discussion.

The quality of articles included in this review was very low and the focus was largely on collections of studies instead of systematic reviews. The median AMSTAR rating was 2/9, which would suggest very low confidence in the conclusions and recommendations of this review. However, as the purpose of this review was to explore the terminology and definitions of the field, and not the conclusions or findings made by any particular articles, this poor quality of included studies should not have an effect on

the quality of this review. Rather, such consistently low quality of research into the field should instead suggest that more high-quality research be conducted to verify or disprove the phenomenon of evidence reversal.

While the quality of included studies may not necessarily affect the quality of the conclusions that this review makes, one limitation that must be considered lies with the definitions of some of the included terms. Many of the included reviews used multiple terms but did not provide all of the corresponding definitions. As a result, our table of terms and definitions has much overlap between the definitions given for different terms from the same article (e.g. Niven *et al.* only provided a definition for "de-adoption," though they found over 40 terms related to the process, many of which we considered to be unique and were separated into different term sets).[97] This limits the strength of our results because some of the sets of terms that we found do not have distinct definitions, despite the terms themselves being unique – even to the point of having different relationships to reversal (e.g. "Re-appraisal" and "Contradict*" which are respectively terms for a consequence and the phenomenon of reversal, were both found in the review article by Niven *et al.*, but neither was defined in its own right, so both are assigned the only relevant definition provided by the article: the definition for de-adoption "The discontinuation of a clinical practice after it was previously adopted" [97]).

2.4 CONCLUSION

Evidence reversal, though not a new phenomenon, has only recently been named. There are many different terms for the phenomenon of reversal as well as the consequences and process that follow identifying practices that are targeted for reversal. Given the similarities between definitions for the various terms, the best way to proceed

with discussion of evidence reversal is under a simplified and all-encompassing definition. Consensus should be reached on which terms are most appropriate so that subject headings can be developed and cohesion can be brought to this emerging field of meta-research. Once there is unity in the theoretical aspects of ER, then researchers can start to investigate more tangible aspects including identifying the targets of reversal and characteristics of initial evidence that may be potential predictors of future reversal.

CHAPTER 3

An exploration of characteristics associated with Evidence Reversal: Methods – Part I

Rationale & Methods of Database Generation

Riaz Qureshi

Chapter Summary: This chapter presents the methods used to construct the database upon which we conduct all of our exploratory analyses. The rationale for our approach is presented first, followed by explanations of our screening and data-extraction methods. More detailed protocols with exact descriptions of the decision-making processes at each stage are presented in Appendices D and E.

CHAPTER 3

3.0 A QUANTITATIVE APPROACH TO REVERSAL

Evidence reversal is the phenomenon whereby new and better evidence contradicts a previously held belief about an established practice or standard of care that was based on weaker evidence. In its current state, the meta-research surrounding the field of medical reversals and evidence reversals remains disorganized and unnecessarily complicated. Since the field is new, there are many different terms and areas of research being explored. In our systematic overview of the literature, we found 87 articles that could be characterized into four broader categories of research surrounding reversal: research about the instance of reversal or contradictory findings, research about the consequences or recommendations for practice change after something has been reversed, research about the practices that are low value and should be targeted as reversal, and research about the characteristics of practices and early research that increases the likelihood of later reversal.[122]

Between these four facets of reversal research we found 50 unique sets of terms that have been used in the literature. The phenomenon of reversal was most often denoted as "medical reversal" and "contradicted / contradict* / refute / contradictory result" while "disinvestment" and "overtreatment, medical overuse, 'too much medicine,' overuse, or misuse" were primarily used to describe the consequences of reversal.[122] We found that the practices targeted as reversals were most often referred to as "low-value [practices / healthcare / services / interventions]" or "[research updates most / trials / POEMS] likely to change clinical practice," and that discussion of potential predictors of future reversal included research characteristics such as "discrepancy," "inconsistency," "uncertainty,"

"known uncertainty," "false positive or inflated results," "initially stronger effects," and "non-replication."[122]

Despite being a quickly expanding field, the vast majority of research pertaining to evidence reversal has been qualitative: we found that the most common type of review article discussing an aspect of reversal was "collection of studies," which were primarily narrative and included letters to the editor, editorials, and recommendations for clinical practice.[122]

3.1 THE CAUSES AND CHARACTERISTICS OF REVERSAL

The phenomenon of evidence reversal is complex and there are two different directions from which the problems posed by unnecessary reversals can be approached. One approach is an analysis of the characteristics of original research that leads to the premature adoption of practices that will later be reversed. Such an analysis would provide insight into the causes of reversal and could lead to the generation of a predictive model for the likelihood of reversal in the future. The other approach would be an analysis of the characteristics of trials that find contradictory evidence against established practices and beliefs. Such an analysis would not provide a predictive model, but could elucidate the characteristics of research that are associated with reversing previous evidence.

In his paper about the high prevalence of research findings that are false or inaccurate, Ioannidis touched on several important characteristics of research that could contribute to future reversal – in relation to the first approach – including: increased financial interests or prejudices, the non-declaration of conflicts of interest, and the novelty of a research field.[23,39] Beyond these contributing factors, other characteristics of

research that may be associated with risk of reversal include: the use and support of traditional practices that have never been tested using randomized controlled trials, the overgeneralization of pathophysiological models to untested indications and populations, and implementing interventions and practices based on their effects on surrogate outcomes.[1,22,30,33,35,36] Furthermore are the common problems with study design and conduct that can affect the validity and fragility of research findings which include, among others: small sample size, low numbers of events, use of restricted populations, biased data collection and assessment, as well as the validity of measures used to assess an outcome. All of these characteristics may be applicable to evidence reversal in assessing the risk of future reversal based on initial studies.

The second approach is also important as the analysis and understanding of the characteristics of contradictory evidence is critical for the development of a framework of reversibility. By 'framework of reversibility,' we mean a conceptual framework of study characteristics that are associated with the contradiction of beliefs about current medical practices that can be used to guide future research and practice. In particular, a framework outlining the prevalence of study characteristics and degrees of association that trials may have with evidence reversal would be valuable to researchers and policy makers in guiding trial design to test current standards of care and the adoption or de-adoption of practices.

3.2 A DATABASE OF "REVERSALS" AND "CONFIRMATIONS"

Prasad *et al.*'s 2013 study, 'A decade of reversal,' was a review of all original articles published in the New England Journal of Medicine between 2001 and 2010.[4] After screening 2044 articles, they determined that approximately 65% (1344/2044)

concerned a medical practice, 27% (363/1344) of which tested an established practice, and of those, 40% (146/363) were declared medical reversals: finding the practice no better or worse than originally believed.[4]

While this was the first systematic attempt to quantify reversal, the study had several limitations, particularly with respect to how reversals were characterized. One element that was not included was the quality of studies. A study was classified as a medical reversal if the original study authors declared that their findings contradicted current practice, and all original research articles, regardless of the study design, were included in their review. However, the element of study quality is inherent to the definition of reversal (i.e. a study must be better quality than its predecessor to reverse that previously established practice). We sought to improve upon this limitation by including only randomized controlled trials, which are assumed to represent higher quality of evidence than other study designs for testing interventions.[123] However, our classification of reversal or reaffirmation was still largely based on what was presented by the authors of each respective RCT included in our database.

Another limitation of 'A decade of reversal' was that the statistical analyses were primarily descriptive of the sample of studies that they had collected: percentages of trials that examined new versus existing medical practices, the distribution of study designs, the percentages of trials that had conclusions that were positive versus negative or no difference between comparators, and the prevalence of studies that were classified as: 'reversal,' 'confirmation,' 'replacement,' or 'back to the drawing board'.[4] Prasad *et al.* also qualitatively described each study that they considered to be reversals. However, no analyses of association of characteristics with reversals were conducted.

Table 3.1 outlines the features of 'A decade of reversal' and directly compares them to the approach taken in this thesis.

Table 3.1 Comparing our approach to Prasad *et al.*'s 'A decade of reversal'

'A decade of reversal'	Expanded study outlined in this thesis

• NEJM (2001-2010)
• All original research studies
• Includes studies of both new and established standards
• Descriptive statistics (study design)
• Qualitative descriptions of 'reversals'
	• NEJM (2000-2016)
	• Randomized Controlled Trials (subset of all original research studies)
	• Includes only trials testing established practices
	• Descriptive statistics (study design)
	• Descriptive statistics (study results, methodology, and quality)
	• Exploratory analyses of association of trial characteristics with reversal of evidence using 3 logistic regression approaches
	  • Univariable
	  • Multi-variable
	  • Backwards-stepwise model selection

To construct our database of reversals and reaffirmations, we first collected all of the same characteristics assessed by Prasad *et al.* To expand upon their analyses and to further explore the characteristics of study design that may be associated with reversal, we also conducted an extensive quality assessment for each included trial using several different approaches. The inclusion of each of the individual components of these assessments, as well as their overall judgements allows us to quantitatively explore the characteristics of reversal to a greater degree than possible in 'A decade of reversal.'

3.3 SCREENING

We collected and screened all articles published as original research studies in the New England Journal of Medicine (NEJM) from January 2000 until December 2016. The NEJM was selected because it was the most cited journal in the medical sciences at the time: based on the 5-year Hirsch Index for Medical Journals.[124] The use of a single

journal was also necessary to restrict the project to a manageable size, given the time constraints of a two-year program.

Collected articles were screened in three consecutive stages on the basis of published abstracts and full-text articles. The first two levels of screening were not conducted in duplicate, but the third level was screened in duplicate (RQ, DS).

In order to be included in this review, articles must have met three criteria: 1) they must evaluate a medical practice; 2) they must be a randomized controlled trial; and 3) they must evaluate an established practice or current standard of care.

As we planned to analyse the characteristics of studies that are likely to be associated with reversal, it necessitated two further inclusion criteria beyond the collection of all studies of medical practices: that of RCT study design – the established gold standard for testing the effect of interventions and consequently assumed higher quality than observational study designs – and that of established practices so as to create a dichotomous outcome upon which to build a logistic regression: contradiction (i.e. reversal) or confirmation (i.e. reaffirmation) of the current practice.

There were no restrictions placed on medical field or setting: all articles that met the above three criteria were included, regardless of their domain. For full descriptions of how decisions were made for each of the three inclusion criteria, including article excerpts to support the description of methods, please see APPENDIX D.

3.3.1 Medical practice

As per the methods described by Prasad *et al.*, articles that tested a medical practice were defined as "any investigation that assesses a screening, stratifying, or diagnostic test, a medication, a procedure or surgery, or any change in health care provision systems".[4]

Besides medical practices, there were several other subject areas that appeared in the NEJM including research articles pertaining to molecular basis of disease, pathophysiology of disease, and animal studies. Articles addressing these subjects were excluded as they did not fit the pre-specified definition of medical practice.

3.3.2 Randomized controlled trial

On the basis of their abstracts and methods, the study design of an article was classified as randomized controlled trial (RCT), prospective controlled (but non-randomized) intervention study, observational study (prospective or retrospective), case-control study, or other methods (including reviews, case series, and case studies). Only RCTs were included in this review; all other study designs were excluded in an effort to create a database of higher quality studies.

3.3.3 Current standard of care or existing practice

The classification of whether or not a trial tested an existing practice was made on the basis of the abstract, introduction, and discussion sections of the papers. While no literature searches were conducted to verify each practice as being currently in use, as this would have been infeasible given the number and extent of searches that would have been necessary, this inclusion criterion was screened by two authors (RQ, DS) who had access to practicing health care providers, and all disagreements were resolved through consensus, and when necessary by consultation with a health care provider. We believe this to be a fair replication of Prasad *et al.*'s methods for determining whether a practice was new or existing, though they were not explicit in describing the criteria used to determine whether a practice was new or existing.

Some trial authors were clear in their description of a practice's prior use, while some chose to downplay or overemphasize the use of a practice. As such, while we felt

that the majority of trials were correctly designated as testing new or existing practices, there were some trials that were contentious and required discussion between RQ, DS, and JM.

3.4 DATA EXTRACTION

As we sought to explore the characteristics of studies that may be associated with reversal, our extraction included any characteristics that we believed may have some relevance to the phenomenon. Thus, the extraction for each RCT included: general identifying information, study design and methodology, study results, overall conclusions, conflicts of interest, PICOTS assessment, Risk of Bias assessment, and overall GRADE rating. Table 3.2 presents each of the characteristics included in the database.

Table 3.2 Database characteristics extracted and automatically completed for each included trial

Extraction Section	Characteristics Extracted	Automated Characteristics
General information	• Authors • Title • DOI • Date of publication • Registered/protocol published • Registration number/protocol citation • Year of trial initiation • Year of trial registration • Year of trial completion	• Year of publication • Years between start and registration • Years between registration and publication • Years between completion and publication • Years since publication
Study design and methodology	• Population • Intervention • Comparison • Primary outcome • Primary outcome: favourable/unfavourable? • Secondary outcomes • Duration of follow up • Sample size • Required sample size • Delta used to calculate sample size • Whether each of the above elements matches the protocol	None

Study results	• Total loss to follow up • Loss to follow up in intervention group • Loss to follow up in comparison group • P-value for primary outcome • Statistical significance of P-value • Point estimate for effect measure • Confidence interval around point estimate • Measure of effect • Type of outcome • Events in intervention group (dichotomous) • Number of subjects in intervention group • Events in comparison group (dichotomous) • Number of subjects in comparison group • Intervention group mean (continuous) • Intervention group standard deviation (continuous) • Comparison group mean (continuous) • Comparison group standard deviation (continuous)	• Percent of sample size lost to follow up • Intervention group rate and confidence interval (dichotomous) • Control group rate and confidence interval (dichotomous) • Absolute risk difference and confidence interval (dichotomous) • Number needed to treat and confidence interval (dichotomous) • Total number of events (dichotomous) • Relative risk reduction (dichotomous) • Fragility index (dichotomous) • Standardized effect size and confidence interval (continuous and dichotomous) • Adequacy of power (continuous and dichotomous)
Overall conclusions	• End point conclusions • Justification for conclusion • Does the article contradict current medical practice? • Justification for contradiction or confirmation • Primary outcome reported in abstract conclusion • Abstract conclusion based on subgroup analyses • Abstract conclusion based on secondary outcomes • If reversal: what category? • Personal judgement on whether trial is a reversal	• Is the article a reversal or a reaffirmation?
Conflicts of interest	• Funding designation • Sources of funding	None
PICOTS assessment	• Sufficiency or Insufficiency of each of the following characteristics o Population o Sample size o Intervention o Comparison o Outcomes o Type of outcome (hard, surrogate, composite) o Follow up o Study design o Study purpose/question	None

	• Justification for the designation applied to each of the above characteristics • Overall sufficiency of PICOTS	
ROB assessment	• The likelihood of risk of bias in each of the following characteristics o Sequence generation o Allocation concealment o Blinding o Incomplete outcome data o Selective outcome reporting o Other design areas • Justification for the designation applied to each of the above characteristics • Overall likelihood of risk of bias	None
GRADE assessment	None	• General risk of bias (Overall likelihood of risk of bias) • Directness and applicability (Overall sufficiency of PICOTS) • Imprecision of results (Adequacy of power) • Modified risk of publication bias (Selective outcome reporting bias) • Total number of downgrades • Overall quality of evidence

Data extraction was completed by three extractors: RQ (years: 2000, 2001, 2002, 2003, 2004, 2006, 2008, 2010, 2012, 2014, 2016), DS (years: 2005, 2007, 2009, 2011, 2013, 2015), and Dr. Leonardo Guizzetti, who was also trained and contributed to data extraction for 2006.

To guide the data extraction process, a protocol for this review, outlined by Sutton *et al.* was followed.[45] This protocol outlined the components and processes followed for each characteristic in the database. The use of the protocol and a random test-set of trials before completing extraction minimized potential differences between extractors. The full table of data-extraction and analysis elements for the database can be found in APPENDIX E.

3.4.1 General information

General information collected from each study was primarily used for identification, although characteristics such as the dates of publication, trial registration, initiation, and completion were also used to inform the PICOTS assessment.

3.4.2 Study design and methodology

Characteristics of the study design and methodology that were extracted include: population, intervention, comparison, primary outcome, secondary outcomes, duration of follow-up in weeks, actual sample size, and the required sample size to meet author specified power level, significance level, and specified differences between point estimates of measures of effect. Furthermore, each of these characteristics was compared with the trial protocol or registration (if available) to inform the PICOTS assessment.

It is important to address the selection of intervention, comparison, and primary outcome because they formed the basis for the characteristics of the study results. All three of these characteristics were attributed based on the authors' designation. In some cases, the designation was not explicit or multiple options were available, in which case pre-specified rules were followed, as outlined in the data extraction protocol (APPENDIX E).

3.4.3 Study results and overall conclusions

Characteristics of the study results that were extracted include: loss to follow up, significance, point estimate and confidence intervals of measures of effect for the primary outcome (if provided), as well as the raw findings pertaining to the primary outcome: numbers of subjects in each group and numbers of events, for dichotomous outcomes, and means/medians and standard deviations/interquartile ranges for continuous outcomes.

In addition to the extracted characteristics, the Excel database was encoded to automatically calculate clinical measures including: point estimates and confidence intervals for the absolute risk difference (ARD), number needed to treat (NNT), the relative risk reduction (RRR), and a standardized effect size for dichotomous and continuous outcomes. The Fragility Index and reverse Fragility Index were respectively calculated for each eligible trial using a web application and the R – 3.3.3 'Fragility Index' package.

The overall conclusions of the included trials were taken from the discussion or the conclusion of the abstract and pertain to the main finding for the primary outcome. Similarly, the classification of the trial as contradicting current medical practice was taken from the authors' conclusions, recommendations, and background in describing the current beliefs surrounding the practice in question.

If a trial was determined to contradict a previously established and currently used practice, the type of reversal was specified as one of several predetermined categories including when the practice was found to be harmful; not effective; less effective than currently believed, but still beneficial; or beneficial if thought to be harmful/not effective/inferior to a different practice.

3.4.4 Conflicts of interest

The conflicts of interest included all sources of funding reported by the authors. If sources of funding or other potential conflicts of interest were stated, they were classified as non-industry or industry. If any of the sources of funding were from an industry company (e.g. pharmaceutical makers), then the conflicts were classified as industry. However, if a company provided only the intervention and this was declared in the paper (i.e. the authors explicitly stated that the company only provided drugs/devices and not

any funding or further support), then the designation was based on the other sources of funding and potential conflicts.

3.4.5 PICOTS assessment

A PICOTS (population, intervention, comparison, outcome, timing, study design) assessment was conducted to determine the sufficiency of the study methodology that was extracted previously. Each component of trial methodology was classified as sufficient or insufficient with regards to its adequacy for reaching an answer for the primary outcome and overall study question, as well as its similarity to the trial protocol/registration (if available). The general guideline that was followed led to a designation of sufficient as being appropriate if the relevant information was itemized or stated in the article or its protocol. If the information was not present or was inappropriately different from the protocol (when available) then the component was designated as insufficient. An overall assessment of the PICOTS was generated on the basis of the sufficiency of the individual components. APPENDIX E contains detailed instructions to guide the designation of each component as sufficient or insufficient.

3.4.6 Risk of Bias assessment

A risk of bias (ROB) assessment was conducted to determine the overall risk of bias for each trial. The ROB tool was developed from the Cochrane Handbook, Chapter 8 and each item is given a designation on a 4-point scale: 'definitely low risk of bias,' 'probably low risk of bias,' 'probably high risk of bias,' or 'definitely high risk of bias.'[125] The ROB assessment is similar to a PICOTS in that it requires one to judge a study's design, however it covers different aspects of methodology where biases may be introduced including: treatment sequence generation, allocation concealment, blinding to intervention groups, handling of incomplete outcome data, whether outcomes were

selected and switched or pre-specified, and other general issues with study design (e.g. early termination, industry influence, extreme baseline imbalance).

The general guideline that was followed for the ROB assessment was that if direct evidence to inform the decision was provided in the article, then ratings of "definitely low risk of bias" or "definitely high risk of bias" were appropriate options. However, if the relevant information was not explicitly reported in the article but could be inferred, or there was insufficient information to permit judgement, then ratings of "probably low risk of bias" or "probably high risk of bias" were appropriate options. A rating of "definitely high risk of bias" was assigned to a domain when there was direct evidence that bias could have been introduced in that design element. An overall assessment of the ROB was generated on the basis of the individual domains. APPENDIX E contains detailed instructions to guide the rating of each component's risk of bias.

3.4.7 GRADE assessment

Grading of Recommendations Assessment, Development and Evaluation (GRADE) is a well-established tool to help determine the overall quality of evidence that a study provides.[126] There are five domains including general risk of bias, directness and applicability, imprecision, risk of publication bias, and inconsistency of findings.[126] All studies included in our review had an initial GRADE rating of 'high quality evidence' because under the GRADE framework, trials are considered to start at the highest level of evidence as opposed to observational studies, which begin at a low level of evidence.[123]

The GRADE assessment in the database was coded to automatically complete for each trial based on the previously extracted characteristics. The GRADE 'risk of bias' was taken from the overall ROB assessment. 'Directness and applicability' was autocompleted with the overall PICOTS assessment. 'Imprecision' was automatically

characterized as 'sufficient' or 'insufficient' based on the adequacy of the trial's power. A modified 'risk of publication bias' was based on the risk of bias for outcomes selection (within the ROB assessment) using reporting bias as a proxy for publication bias.

We did not include the GRADE domain of 'inconsistency' in our assessment because it is specifically used for describing the heterogeneity of results across multiple trials on the same topic. As we collected all original RCTs of medical practices, with no restriction on type or field of practice, it was impossible to describe the degree of consistency across the evidence. Therefore, we made the simplifying decision to not assess the inconsistency of included trials.

Another simplifying assumption that we made in our GRADE assessment was the inability to increase quality of evidence rating through inflating factors. The GRADE framework allows studies to increase their quality rating if they exhibit: a large or very large effect size, a dose-response relationship, or if the presence of residual confounding would reduce the demonstrated effect or suggest a spurious effect if no effect was observed.[127] However, this upgrading of the evidence is primarily only applied to observational studies and though it is theoretically possible to upgrade the quality rating of RCTs, the GRADE Working Group remark that they "have yet to find a compelling example of such an instance."[127] Also, by including only RCTs, each article was automatically categorized as the highest quality of evidence initially. For these reason, we decided that for our GRADE assessment, trials – and consequently all included studies in this review – would not be able to receive any upgrading of evidence as proposed within standard GRADE methodology.

3.5 CONCLUSIONS

Prasad *et al.*'s review from 2013 provided a starting point from which we sought to further explore the characteristics of research that may be associated with reversal. To this end, we independently replicated and also further expanded their methodology to allow for more in-depth analyses of quality, methodology, and the findings of RCTs that test established practices and may lead to evidence reversal.

CHAPTER 4

An exploration of characteristics associated with Evidence Reversal: Methods – Part II

Data Analysis Plan

Riaz Qureshi

Chapter Summary: This chapter presents the methods used for all of our planned analyses and additionally serves as a protocol to guide the analyses as it was written before the analyses were conducted. The planned reproduction and expansion of 'A decade of reversal' is presented first, followed by an explanation of our exploratory logistic regression analyses. The Stata do-files for setting up the database and conducting the analyses are presented in Appendices F and G.

CHAPTER 4

4.0 STATISTICAL ANALYSES PLAN

The 'Decade of Reversal' study by Prasad *et al.* was a landmark review in the field of medical and evidence reversal as it was the largest and most comprehensive study to specifically address the phenomenon.[4] Given its importance, and their focus on the years 2001 to 2010, an independent replication of the study, together with a further expansion to include more recent articles (2000 to 2016), is needed to assess reproducibility and to add power. Furthermore, reproducing the review provides an opportunity to expand the breadth of analyses to identify trial characteristics that may be associated with reversal, potentially providing the necessary data to create an evidence-informed framework of reversibility to guide future research.

After screening all articles published in the New England Journal of Medicine between January 2000 and December 2016, the characteristics of included articles were extracted into an excel database as described in Chapter 3 and APPENDICES D and E. All descriptive statistics were calculated using Stata 13, as were all regression analyses. The Stata do-files for importing and setting up the excel database and then conducting the analyses described in this chapter can be found in APPENDIX F and APPENDIX G, respectively.

4.1 ANALYSES IN 'A DECADE OF REVERSAL'

The 2013 review was conducted to identify medical practices that offer no net benefits. The authors reviewed all articles published in the NEJM between 2001 and 2010, and classified them according to whether they addressed a clinical practice, whether they tested a new therapy or an existing therapy, whether the final results and

conclusions were positive versus negative or no difference (i.e. whether they found a significant effect favouring their intervention or control group, or found the groups were not statistically different), and whether the results constituted evidence "replacement," "back-to-the-drawing-board," "reaffirmation," or "reversal."[4]

The analyses that Prasad *et al.* conducted in 'A decade of reversal' were primarily descriptive. In their results they described: the proportion of articles addressing a medical practice (65.8%); the proportion of medical practices that were new versus existing (73% and 27%, respectively); the proportions of different study designs (67.7% RCTs, 16.4% prospective controlled but non-randomized studies, 8.7% observational studies, 3.2% case-control studies, and 3.9% studies with other methods); the proportions of studies reaching conclusions that were positive (significant difference in favour of intervention) versus negative (significant difference in favour of control or non-significant difference) between comparators (70.5% and 29.5%, respectively); as well as the overall proportions with conclusions that constituted replacement, back to the drawing board, reaffirmation, or reversal (56.3%, 12.3%, 10.9%, and 10.3%, respectively).[4]

In addition to the above descriptive statistics, Prasad *et al.* specified the most common study type, the proportions of reaffirmations and reversals among studies that tested existing medical practices, the proportions of study types among articles that constituted reversals, and the statistical likelihood that articles testing new or existing practices would find the practice to be beneficial or ineffective. The remainder of the results presented in the review were qualitative descriptions of selected reversals and a limited exploration of their trends; namely the narrative shared by many reversals, which entails the acceptance of a practice or standard – despite a weak evidence base – due to

support from prominent members of the medical community and a faith in the pathophysiologic mechanism, which is subsequently undermined when adequately tested by properly conducted randomized controlled trials.[4]

4.2 INDEPENDENTLY REPRODUCING AND EXPANDING ANALYSES OF 'A DECADE OF REVERSAL'

In an effort to independently reproduce, expand, and update 'A decade of reversal,' we attempted to classify articles into the same categories over the same years as in Prasad *et al.*, with an additional 7 years of trials. We replicated their methodology as accurately as we could from the description provided in the article. However, we also wanted to improve upon their methods by assessing the characteristics of studies that may be associated with reversal. To this end, we only extracted data for articles that were classified as randomized controlled trials (RCTs) as they are assumed to provide a higher quality of evidence than observational studies, and generally provide conclusions that are more robust.

The baseline analyses of our database of RCTs will include independent reproduction of all descriptive statistics presented in 'A decade of reversal.' This will include: the proportion of articles that address a medical practice, the proportion of medical practices that were new versus existing, the proportions of different study designs testing medical practices, the proportions of article conclusions being positive (significant) versus negative (non-significant difference) between comparators, and the proportions of reaffirmations and reversals among trials that tested established practices.

There are several descriptive statistics that were not conducted because articles that were classified as testing "new" practices were excluded from our study and have no

further data extracted beyond that collected to inform the first two levels of screening (outlined in Chapter 3, Section 3.3). Further, the "study type" of all included articles for our study is "RCT," and thus descriptive statistics relevant to other study designs do not apply to our study sample. We will not describe: the overall proportions of studies designated as "replacement" or "back to the drawing board," the most common study type, the proportions of study types among articles that constituted reversals, and the statistical likelihood that a trial testing new practices would find the practice to be beneficial or ineffective.

4.3 EXPANDED ANALYSES: DESCRIPTIVE STATISTICS

Prasad *et al.* explored the presence of reversal within the NEJM and described several characteristics of the articles that they found. Our goal was to expand the analyses conducted in 'A decade of reversal.' Consequently, we collected additional characteristics about the included trials, which allowed us to provide improved description of the sample of trials and characteristics that may be associated with reversal. In addition to independently reproducing the descriptive statistics presented by Prasad *et al.*, as outlined above in section 4.2, we also report other descriptive statistics for our sample as would be found in an observational study or trial. These will be reported for the overall sample and according to whether the trial contradicted or supported the practice that it was examining (i.e. reversal vs. reaffirmation).

In our tables of sample characteristics, we will report: proportion of trials that were registered; proportion of those registered that had an accessible protocol or registration; mean number of years between trial start and registration of those where the protocols/registrations were accessible; the proportion of trials that had a primary

outcome that was focused on harm; mean duration of follow up; mean sample size achieved; mean required sample size; mean percentage lost to follow up; proportion that had significant primary outcomes (with a P-value ≤ 0.05); proportions of studies using different measures of effect; proportion with a primary outcome that was dichotomous; proportions of studies with primary outcomes that are based on hard, composite, or surrogate outcomes; proportions of studies that reported abstract conclusions based on their primary outcome, subgroup analyses, or secondary outcomes; proportions of studies belonging to each category of 'reason for reversal;' proportions of studies with each overall PICOTS designation; proportions of studies with each overall ROB designation; and proportions of studies with each overall GRADE level of evidence.

4.4 EXPANDED ANALYSES: LOGISTIC REGRESSION

While descriptive statistics are valuable in characterizing a sample of studies that may lead to evidence reversal, the conclusions that can be drawn from such analyses are limited. A multivariable logistic regression of the characteristics of these studies would provide a greater understanding of the degree to which they are associated with reversal. Logistic regression will be used for this analysis because it is the most widely used model for binary outcomes in clinical and epidemiological applications.[128] This quantitative approach to assessing the characteristics of trials will contribute to the generation of a framework of reversibility to help guide future research in the field.

The multivariable logistic regression model assumes that multiple covariates are related to the outcome in an additive fashion on the log scale.[128] Other assumptions about the covariates and outcome include: the model is fitted correctly; the outcome follows a binomial distribution; the mean expected outcome for a given set of covariates ($E[y|x]$ =

P(x)) is given by the logistic function; values of the outcome are statistically independent (i.e. truly binary); all observations are independent; and the requirement of large sample sizes.

To ensure the model has appropriate fit, Pearson and Hosmer-Lemeshow (HL) goodness-of-fit tests will be performed. If either goodness-of-fit test suggests that the model is not appropriate (P-value ≤ 0.05) then the least non-significant covariate will be removed and the goodness-of-fit retested, until the model is appropriate. The outcome of the model (reversal versus reaffirmation) is binary and the requirement of large sample sizes will be assumed met for all univariable analyses, as there will be more than 10 reversals (cases) per covariate given the sample size of 611 trials. However, there will be reduced power for the overall multivariable regression as controlling for several categorical covariates will lower the number of trials with each specific designation. An assumption that may not be met is the independence of all observations, as some trials included in the database are related. As all RCTs from 2000-2016 that met the inclusion criteria are in the database, some included trials are secondary analyses of earlier trials (which may also be included), or multiple publications on different outcomes of the same trial (e.g. publication of intermediary analyses and final outcome data). However, the simplifying assumption of independence between observations will be applied as only a small proportion of the 611 included trials are not independent.

This project presents a comprehensive study of RCTs published in the NEJM, but it is neither a meta-analysis, nor a meta-regression. Rather, the sample of trials that have been collected will be treated as 'individuals' in assessing the relation of their characteristics to a known outcome. In this sense, an analogous study design to the

overall approach could be described as a repeated cross-sectional study. A case-control design would not be an appropriate analogue, as a sample of reversals and reaffirmations was not selected at the start of the trial based on their outcome. Neither would a cohort design be an appropriate analogue because there are different 'subjects' each year and there are no repeated measures of the same 'individuals.' A repeated cross-sectional study design is the most appropriate as the database is a sample of trials each year for the past 17 years – selected on the basis of inclusion criteria that were designed to refine the sample to one most likely to have the outcome of interest – and extracted relevant characteristics for which the relation to the outcome of interest (determined after inclusion in the study) will be described. The implication of this 'approach' is that we cannot infer causality or influence of the characteristics on the outcome, only the degree to which they are associated, because all of the data for the characteristics and the outcome is collected at the same time.

4.4.1 Overall logistic regression

In order to assess the strengths of associations that the characteristics of trials may have with reversal, potentially important characteristics will be included as covariates in a multivariable logistic regression on the outcome of reversal (contradiction of established practice) or reaffirmation (confirmation of established practice). Table 4.1 presents the covariates that will be included in the overall logistic regression model, as well as their possible values. Section 4.5 describes the rationale for inclusion of each covariate in the model as well as the methods for describing and assessing their distribution and validity of inclusion in the model.

In the logistic regression analyses of these characteristics, the associations of all individual covariates were first tested in univariable analyses. An overall logistic

regression model was then fitted with all potential predictors in a simple exploratory analysis, which was followed by a backwards-stepwise model selection. As an additional approach to assessing the covariates, we assessed the correlation of each covariate with the others.

Table 4.1 The 15 covariates included in overall logistic regression

Type of Covariate	Covariate name	Possible values
Continuous	1. Percentage of participants lost to follow up	0.0 to 100
	2. Length of follow up	0 to (+ ∞)
	3. P-value	0.0 to 1.0
	4. Sample size	0.0 to (+ ∞)
	5. Standardized effect size	0.0 to (+ ∞)
	6. Year of publication	2000 to 2016
Binary	7. Protocol registered	• Yes • No
	8. Abstract conclusions based on primary outcome	• Yes • No
	9. Abstract conclusions based on secondary outcome	• Yes • No
	10. Abstract conclusions based on subgroup hypotheses	• Yes • No
Categorical (Ordinal)	11. Overall PICOTS assessment	• Sufficient • Somewhat insufficient • Clearly insufficient
	12. Overall ROB assessment	• Definitely low risk of bias • Probably low risk of bias • Probably high risk of bias • Definitely high risk of bias
	13. Overall GRADE assessment	• High • Moderate • Low • Very low
Categorical (Nominal)	14. Conflicts of interest	• Non-industry • Industry • None reported
	15. Type of outcome	• Hard • Composite • Surrogate

The logistic backwards-stepwise regression fits all explanatory variables to a model and then sequentially removes the covariates that are non-significant by a pre-

specified level for removal.[129] The model iteratively checks the significance of the least significant included covariate to see if it is less than the pre-specified level for exclusion from the model and re-estimates the fit after each removal until all include covariates are significant at the pre-specified level.[129] This method theoretically produces the best fitting model from a set of potential predictors, but only when there is no prior subject matter knowledge to enable pre-specification of important covariates.[130]

There are limitations to the use of backwards-stepwise model building – particularly with regards to the stability of the selection process in the presence of collinearity – such as over-fitting the model to the data, yielding highly biased R^2 values and models that are not generalizable.[130,131] However, as this analysis is exploratory, and we wished to determine if any of our pre-determined covariates have relationships with reversal – whether significant or not – and not necessarily to build the most appropriate predictive model, we deemed the backwards-stepwise approach acceptable. Our model selection criteria were lenient and we used a significance level of '0.5' for dropping covariates from the model.[130] Thus, if a covariate has a p-value > 0.5 it may be excluded from the model.[131] The Hosmer-Lemeshow goodness of fit for the original model was compared to the final model generated with the backwards-stepwise method.

The results of the individual covariate analyses and overall logistic regression are presented as a table of the odds ratios, confidence intervals, and p-values for the relationship that each covariate has with the outcome of reversal. These associations between each covariate and the outcome were interpreted from the perspective of trial methodologists to inform the development of a framework which attempts to incorporate

the relationships discovered into a unified decision aid for directing future research and assessing generalizability for detecting past reversals or predicting future reversals.

4.4.2 Logistic regression of multidimensional summary scores

After the general logistic regression had been conducted with all of the covariates of interest, it was possible that one or more of the multidimensional summary components may have been found significantly associated with reversal. The overall GRADE, PICOTS, and ROB assessments are summary scores that are comprised of individual components – covering different aspects of study design and methodology. If any of these summary measures were found to be significantly associated with the contradiction of established practices, we planned to conduct multivariable logistic regressions of the component domains on the outcome to determine which of the components drives the significance of the overall measure.

Table 4.2 outlines each of the three smaller logistic regressions that would have been conducted to assess the individual components of the multidimensional summary scores, if any of them had been significantly associated with the outcome in the overall logistic regression. These models also would have been assessed with Pearson and HL goodness of fit tests.

Table 4.2 Model covariates for each of three separate summary score regressions

Significant summary score	Covariates	Possible values
PICOTS (8 components)	• Population • Sample size • Intervention • Comparison • Outcomes • Follow up • Study design • Study purpose/question	• Sufficient • Insufficient
ROB (6 components)	• Sequence generation • Allocation concealment • Blinding	• Definitely low risk of bias • Probably low risk of bias • Probably high risk of bias

	• Incomplete outcome data • Selective outcome reporting • Other design areas	• Definitely high risk of bias
GRADE (4 components)	• General risk of bias	• Definitely low risk of bias • Probably low risk of bias • Probably high risk of bias • Definitely high risk of bias
	• Directness and applicability	• Sufficient • Somewhat insufficient • Clearly insufficient
	• Imprecision of results	• Sufficient • Insufficient
	• Modified risk of publication bias	• Definitely low risk of bias • Probably low risk of bias • Probably high risk of bias • Definitely high risk of bias

4.5 RATIONALE FOR COVARIATE INCLUSION

The initial set of explanatory variables that were selected for inclusion in the model included a mix of continuous, binary, ordinal, and nominal categorical. There were 15 covariates that we believed might be associated with the outcome of reversal or confirmation of practice because they have been previously identified as indicators of study quality and strength of evidence. Given that a common trend for many reversals involves high quality randomized controlled trials that contradict a practice implemented on a weak evidence base, we assumed that these common markers of study quality may represent good candidate markers with plausible mechanism for relationship with reversal or reaffirmation of established practices.

When there is a high degree of loss-to-follow up within a trial, it can be difficult to interpret the effect of the intervention.[132] Differential or non-random loss to follow up – in terms of numbers and reasons between comparison groups – is particularly challenging as it may affect the validity of trial conclusions, while a low loss-to-follow up can be indicative of good trial design and conduct.[132]

As experience in clinical research has developed, it has become apparent that interventions that were believed to have short-term effects will continue to elicit effects and impact patient outcomes in the long term (e.g. excess mortality among patients with sepsis admitted to the hospital, compared to the general population, remains for several years, yet most interventional sepsis studies use end-points of mortality at 28-days).[133,134] As a consequence, the study of clinical outcomes almost always requires longer lengths of follow up to find the true net effects of an intervention. Thus, duration of follow up is included as a continuous covariate because theoretically, the longer patients are followed, the more likely it is that the true net effect of a practice will be found and potentially reversed.

While sample size is a predictor of significance, both the P-value and sample size were included in the overall model because practices that are established based on small studies may be overturned by large, adequately powered, trials, and thus these characteristics may have some relation to evidence reversal. Similarly, a standardized effect size was included in the model to allow comparison between the various measures of effect, both dichotomous and continuous, used by different trials.

We investigated the year of publication as a potential predictor for two reasons. Firstly, Prasad *et al.* tested if there was a significant relationship between the percent of reversals over time using a linear regression. Even though they found that the percentage of reversals among articles that tested a standard was consistent across the decade (P = 0.51), we still included year as a covariate because, theoretically, the risk of reversals may change as time progresses, and our study has increased the number of years from 10 to 17. This assumption is logical because the longer that a practice has been implemented,

the more opportunities arise where it may be reversed and consequently practices may be significantly more likely to be reversed as time progresses.

The sources for abstract conclusions (primary outcome, secondary outcomes, or subgroup hypothesis) were included in the overall model as dichotomous markers for potential reporting bias. The conclusions of the abstract should report the general interpretation of results and be consistent with the primary outcome reported in the abstract.[135] Publications that make abstract conclusions based on subgroup analyses or secondary outcomes (i.e. selective reporting) may be trying to draw attention away from an unfavourable or insignificant primary outcome.[136]

The overall PICOTS, ROB, and GRADE assessments were also tested in the model because they are summary measures of several multidimensional characteristics concerning study quality. We believe that designations of higher study quality may be associated with reversal because for a practice to be reversed there must be sufficient evidence to support the decision and poor quality studies are less likely to constitute sufficient evidence. This rationale is further supported by the definitions of medical and evidence reversal, which include the qualification that the new evidence claiming to reverse an established practice be superior to that which preceded it.[1]

The designation of potential conflicts of interest and sources of funding were included in the model as trials with industry influence may be more likely to lead to confirmation. This is because all trials in our database test established practices. While industry trials of new practices are more likely to find significant differences, when testing interventions that are already adopted – particularly those created by their own company – they may be more likely to confirm what is believed than contradict it.[137,138]

Outcome type was the final covariate included in the overall model as its relationship with reversal is one of the trends seen among many reversals: practices implemented based on intervention effects on surrogate outcomes are subsequently reversed when the relevant 'hard' (clinically-relevant; patient-important) outcomes are tested.[2,77] We included this characteristic because we expected that trials using surrogate outcomes or composite outcomes may be more likely to confirm the practices that they are testing, while trials investigating an intervention's effect on hard, patient-important, outcomes may be more likely to lead to reversal.[139]

In addition to the 15 included in the overall model, there are five covariates for which the relationship with reversal was only examined in univariable analyses. These are 'years between trial start and registration,' 'years between trial end and publication,' 'Fragility Index,' 'Total Number of Events,' and 'Adequacy of Power.' These covariates are of interest because they are related to the confidence that can be expressed in a trial's results, but must be assessed on their own because all but the Adequacy of Power have high degrees of missing data that cannot be meaningfully imputed, and the Adequacy of Power is an indicator variable of our devising that we do not feel comfortable influencing the potential relationships of other covariates.

The two continuous covariates that measure the years between start and registration, and completion and publication, are indicators of reporting and publication biases as trials with large positive values indicate retroactive registration and long periods of non-publication. These two covariates were not included in the overall model because they only had numerical values when a registered protocol is accessible and have a designation of 'N/A' when there is no registration available. Consideration was given to

including them in the model with a binary indicator variable (the availability of a protocol or registration), but after deliberation with committee members, the decision was made to explore them on their own. The Fragility Index is an indication of how many events would be required to change the significance of a trial's results.[140] A lower Fragility Index suggests that a trial's conclusions may have been different with a few more events in one group or the other and, consequently, that the results may be 'fragile' and more easily reversed.[140] A Fragility Index value of '0' may arise because of a difference in the statistical test used to determine significance, as the Fragility Index calculates p-values using the Fischer's Exact Test.[140] The total number of events is closely linked to the Fragility Index as it is calculated using the number of events and subjects in each group. However, investigating the relationship of total number of events is also of interest because it is a more familiar metric to the medical community and an established contributing factor to the power of a trial in making conclusions. The adequacy of power is a dichotomous indicator of whether or not a trial had sufficient power based on its actual sample size, its reported necessary sample size, and the desired delta between comparators (if reported).

4.6 MISSING DATA

The NEJM was selected as the journal upon which to conduct this review based on its 5-year impact factor (Hirsch-index).[4] The NEJM is widely regarded as being one of the highest quality medical journals in the world and as such, maintains a high standard in the reporting and writing of the articles it publishes. However, even within this high impact journal, some elements of trial design and results were missing from the descriptions provided in the publication (as well as provided appendices and protocols).

In an effort to account for missing data, simple mean imputation was used for characteristics where appropriate. Each imputed characteristic underwent sensitivity analyses to test whether use of the imputed data significantly affected the resulting relationship. The amount of missing data for each covariate – before and after imputation – will be presented in Chapter 5 as the number of observations contributing to the result for each characteristic.

Table 4.3 outlines the methods used for imputation of missing data for each of the covariates in the general model. There are several covariates for which there was no missing data, including: year of publication, whether the trial was registered, the end point conclusions, overall PICOTS assessment, overall ROB assessment, overall GRADE assessment, the reason for reversal, the designation of conflicts of interest, and the type of primary outcome. These are mostly covariates that were our judgements and interpretations of aspects of the trial – based on what the author had presented in their paper – and therefore cannot be missing because they are not directly taken from the publication.

Table 4.3 The 15 covariates and proposed imputation methods for missing data

Covariate	Proposed method of imputation
% Subjects lost to follow up	Mean imputation with average % subjects lost to follow up
P-value	If raw data or effect measure for primary outcome is provided, the missing p-value will be imputed with a mean significant p-value for significant trials with a significant outcome confidence interval, and the mean non-significant p-value for trials with non-significant confidence intervals for their primary outcome
Standardized effect size	Standardized effect size is calculated using the raw trial data. For trials with dichotomous outcomes, the number of subjects and number of events for each group are used to calculate an absolute risk difference, which is used to generate a standard effect size. For trials with continuous outcomes, the mean and standard deviations for each group are used to generate a standard effect size. For trials with no raw data, missing effect sizes will be mean imputed.
Length of follow up	Mean imputation with average duration of follow up

Year of publication	No missing data
Abstract conclusion primary	No missing data
Abstract conclusion secondary	No missing data
Abstract conclusion subgroup	No missing data
Protocol registered	No missing data
End point conclusions	No missing data
Overall PICOTS assessment	No missing data
Overall ROB assessment	No missing data
Overall GRADE assessment	No missing data
Conflicts of interest	No missing data
Type of outcome	No missing data

4.7 DEVELOPING A FRAMEWORK OF REVERSIBILITY

An initial framework of reversibility has been proposed by Sutton and Martin that focuses on specific components of the design, execution, and analysis of evidence by using indicators derived from trial design, methodology, and reporting.[45] The framework – which includes the individual domains of a PICOTS, ROB, and modified GRADE assessment, as well as modified optimum information size, fragility index, duration from trial start to registration and from completion to publication, and sources of abstract conclusions – is proposed as a tool to inform the likelihood that a trial reverses an established practice.[45] This framework may aid healthcare decision-makers in delaying the adoption of new practices or disinvesting established practices until the evidence has matured. Our expanded analyses will further the development of this framework as we explore the relationship that these characteristics and others have with the declaration of reversal.

CHAPTER 5

The characteristics of Reversal: Results

Descriptive Statistics and Logistic Regression Analyses

Riaz Qureshi

Chapter Summary: This chapter presents the results of both descriptive analyses (the reproduction and expansion of those conducted in 'A decade of reversal') and the exploratory logistic regression analyses.

CHAPTER 5

5.0 SCREENING

Three thousand five hundred and sixty original research studies published in the New England Journal of Medicine (NEJM) between January 1st, 2000 and December 31st, 2016 were collected by two authors (RQ and DS). These articles were screened at three levels for inclusion criteria using the abstract and full texts, leading to exclusions of: 834 for not studying a medical practice, 964 for not being randomized controlled trials, and 1147 for testing new practices.

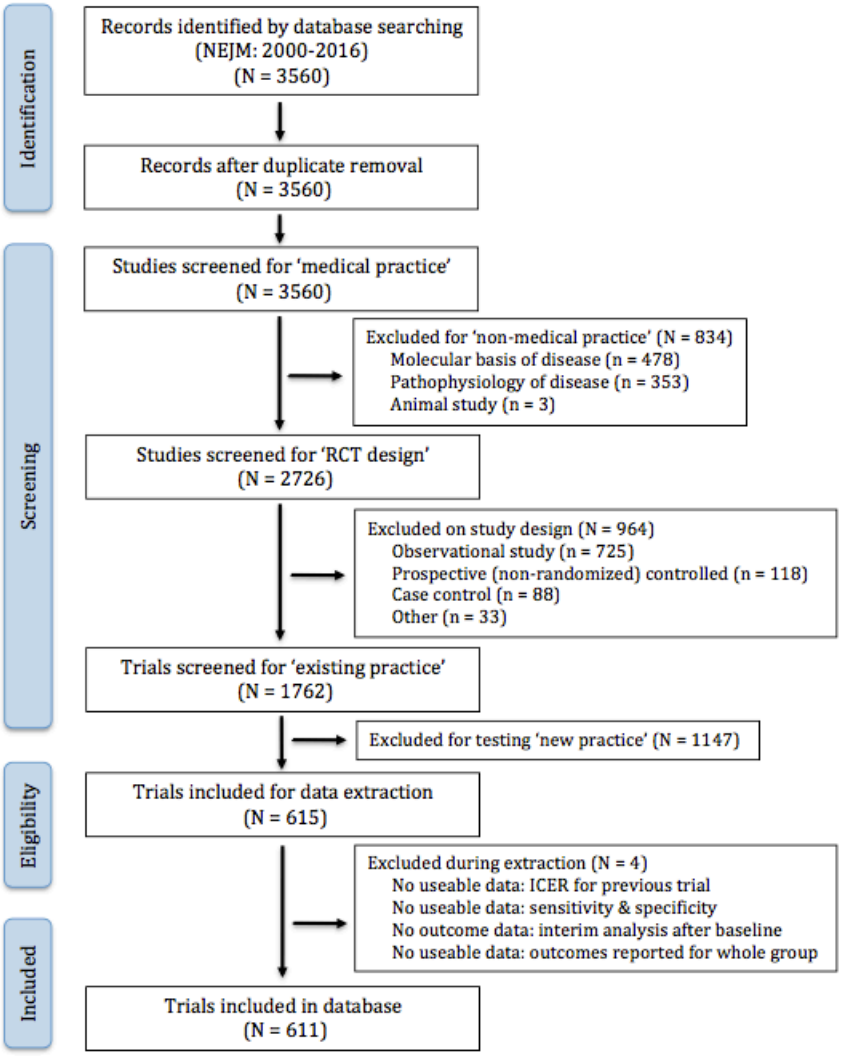


Figure 5.1 PRISMA flow diagram for inclusion of trials

Screening for the first two levels was not conducted in duplicate, but the third tier – existing versus new practices – was duplicated and had an initial Kappa of 69% between the two authors (RQ and DS). After screening for exclusion criteria, 615 trials were included in the database. However, during data extraction, four trials were excluded for reasons that were not accounted for in the initial exclusion criteria. These trials were excluded primarily due to the lack of useable outcome data. One publication was an ICER (Incremental Cost-Effectiveness Ratio) analysis of a previously published trial. One trial was of a screening test, but provided no summary outcome data that could be used. One publication was an interim analysis of the participant's baseline data and had no outcome data for either group. One trial did not present outcome data for each intervention group, but as summary data for the entire study group. Therefore, the database includes 611 randomized controlled trials of established medical practices.

5.1 DESCRIPTIVE STATISTICS FOR INCLUDED TRIALS

Between 2000 and 2016, 3560 original research studies were published in the NEJM. The majority of studies (2726 [77%]) addressed a medical practice. Of the studies concerning a medical practice during these 17 years, we identified 1762 (65%) randomized controlled trials, 725 (27%) observational studies, 118 (4%) prospective controlled but non-randomized studies, 88 (3%) case control studies, and 33 (1%) studies of other design. Of the randomized controlled trials, 615 (35%) were determined to address an existing practice. Characteristics of the 611 included trials were extracted and the descriptive statistics for the sample, presented in Tables 5.1 through 5.4, were calculated using Stata 13.

Table 5.1 Classifying characteristics of studies screened

Screened Studies (2010 – 2017) (n = 3560)	# of Studies (%)	# of Observations
Articles that addressed a medical practice	2726 (77%)	3560
Articles with study design of: Randomized Controlled Trial Observational (prospective or retrospective) Prospective (non-randomized) Controlled Trial Case-Control Other (meta-analysis, case-study, case-series)	 1762 (65%) 725 (27%) 118 (4%) 88 (3%) 33 (1%)	2726
RCTs that addressed an existing practice	615 (35%)	1762

Based on authors' conclusions and the information presented within their backgrounds and discussion, 331 (54%) were reversals – contradicting the established practice being tested – while 280 (46%) confirmed what was believed or upheld the standard of care over a lesser or prior standard (reaffirmation). With regard to the trial results, 256 (42%) reached positive conclusions while 355 (58%) reached negative conclusions or found no statistically significant difference between their comparators.

Among trials that contradicted the established practice being tested, there were several different possible reasons for reversal including: the practice was found to be harmful if it was thought beneficial (19%); the practice was found to be ineffective to the comparator if it was believed to be effective (46%); the practice was found to be less effective or equivalent to a comparator if it was believed to be superior (19%); or the practice was found to be beneficial if it was originally believed to be harmful, not-effective, or inferior to another practice (16%).

Table 5.2 Primary descriptive statistics characterizing evidence reversal

Characteristics of included Randomized Controlled Trials (n = 611)	# of Trials (%)	# of Observations
Authors declarations regarding the tested practice: Contradiction (evidence reversal) Confirmation (evidence reaffirmation)	 331 (54%) 280 (46%)	611
Trial conclusions regarding primary outcome: Positive Negative or no difference	 256 (42%) 355 (58%)	611

Reason for contradiction of established practice:		
Not effective if thought effective	152 (46%)	
Less effective if thought beneficial	64 (19%)	331
Harmful if thought beneficial	63 (19%)	
Beneficial if thought harmful/not-effective/inferior	52 (16%)	

Additional characteristics describing the methodology, findings are described in Table 5.3. Concerning registration, 89% of trials were registered, and of those, 86% were accessible and the average duration between trial start and registration was 1.20 years. Four hundred forty four (73%) trials had a primary outcome that was oriented around harm and 276 (45%) trials had a primary outcome that was significant with a P-value ≤ 0.05. Among included trials: the mean duration of follow up was 115.34 weeks; the mean percentage of subjects lost to follow up was 7%; the mean sample size was 3305; and, if provided by the authors, the mean required sample size was 2184.

The majority of trials had a primary outcome that was dichotomous (474 [78%]), and the results for primary outcomes were presented with a variety of different measures of effect including: Hazard Ratio (27%), Relative Risk (19%), Absolute Risk (17%), Odds Ratio (8%), Effect Size / Mean Difference (11%), Relative Risk Reduction (1%), and 18% where no measure nor magnitude of effect was provided (i.e. Not Applicable). The primary outcomes of trials were most often (45%) based on hard, patient-important, response variables – such as all-cause or cause-specific mortality, risk of stroke, or myocardial infarction – but some studies used composite outcomes (i.e. combinations of outcomes) (31%) or surrogate outcomes (e.g. physiological measures or laboratory values) (23%) as their primary outcome. The abstract conclusions of most trials (520 [85%]) were based on the primary outcome, but 232 (38%) abstracts were based on secondary outcomes and 52 (9%) were based on subgroup analyses. The reason why the

proportions of abstract conclusions do not sum to 100% is because they are not mutually exclusive and may have been derived from the primary outcome, and/or the secondary outcomes, and/or the subgroup analyses.

Table 5.3 Secondary descriptive statistics characterizing included trials

Characteristics of included Randomized Controlled Trials (n = 611)	# of Trials (%) or Mean (Std. Err.)	# of Observations
Trials registered	542 (89%)	611
Protocol / registration accessible	464 (86%)	542
Mean duration between trial start and registration (years)	1.20 (0.13)	412
Trials with an unfavourable primary outcome	443 (73%)	611
Mean duration of follow up (weeks)	115.34 (6.77)	600
Mean sample size	3305 (467.50)	611
Mean required sample size (where provided)	2184 (226.14)	477
Mean loss to follow up as proportion of total sample size	0.07 (0.004)	598
Trials with significant primary outcomes ($P \leq 0.05$)	276 (45%)	611
Trials with a primary outcome measure of effect:		
HR (Hazard Ratio)	166 (27.2%)	
RR (Relative Risk)	115 (18.8%)	
AR (Absolute Risk)	103 (16.9%)	
ES / MD (Effect Size / Mean Difference)	65 (10.6%)	611
OR (Odds Ratio)	49 (8%)	
RRR (Relative Risk Reduction)	5 (0.8%)	
NNT / NNH (Number Needed to Treat / Harm)	0 (0%)	
N/A (Not Available)	108 (17.7%)	
Trials with a dichotomous primary outcome	474 (78%)	611
Trials with a primary outcome that is:		
Hard (i.e. clinical / patient-important)	277 (45%)	611
Composite	192 (31%)	
Surrogate	142 (23%)	
Trials reporting abstract conclusions based on:		
Primary outcome *	520 (85%)	611
Secondary outcome *	232 (38%)	
Subgroup analyses *	52 (9%)	

* Proportions of sources for abstract conclusions do not sum to 100% because abstract conclusions could be derived from none or all three of the sources

In comparing descriptive statistics between trials that were classified as reversals versus reaffirmations (APPENDIX H: Table 6), the two are largely comparable. The most notable difference between the two groups is the proportion of trials having significant findings with regard to their primary outcomes, which was 58% among trials that confirmed the tested practice, but only 34% among trials that contradicted the practice.

The general quality of the trials included in this review was low, as can been seen in Table 5.4 as the greatest proportions of studies had PICOTS, ROB, and GRADE designations of 'Somewhat Insufficient' (48%), 'Probably Low Risk of Bias,' (35%), and 'Very Low Quality' (31%).

Table 5.4 Descriptive statistics for quality assessments of included trials

Quality assessments for included Randomized Controlled Trials (n = 611)	# of Trials (%)	# of Observations
Trials with overall PICOTS designation:		
Sufficient	243 (40%)	611
Somewhat insufficient	294 (48%)	
Clearly insufficient	74 (12%)	
Trials with overall ROB designation:		
Definitely low risk of bias	165 (27%)	
Probably low risk of bias	212 (35%)	611
Probably high risk of bias	167 (27%)	
Definitely high risk of bias	67 (11%)	
Trials with overall GRADE level of evidence:		
High	124 (20%)	
Moderate	168 (28%)	611
Low	128 (21%)	
Very low	191 (31%)	

Figures 5.2, 5.3, and 5.4 depict the distribution of the individual components for the PICOTS, ROB, and GRADE assessments for the whole sample. The distributions of overall assessments for PICOTS were: 40% 'Sufficient,' 48% 'Somewhat insufficient,' and 12% 'Clearly insufficient.' The distributions of overall assessments for ROB were: 27% 'Definitely low,' 35% 'Probably low,' 27% 'Probably high,' and 11% 'Definitely high.' The distributions of overall GRADE quality of evidence scores were: 20% 'High,' 27% 'Moderate,' 21% 'Low,' and 31% 'Very low.'

The PICOTS component that contributed the most to decreasing the sufficiency was the 'Sample size:' 42% of trials had either a sample smaller than required by their reported power calculation or failed to report a required sample size. The ROB component that contributed the most to increasing the likelihood of bias in a trial was

'Other:' 20% of trials had some element of trial design or conduct that implied a high risk of bias and was not captured in the other components (e.g. industry design/conduct/analysis, stopping early for statistical reasons, extreme baseline imbalance, or bias related to the study design/conduct/analysis/reporting). The GRADE component that contributed most to downgrading of evidence was 'Directness and applicability:' 48% of trials were downgraded by -1 for having overall PICOTS of 'Somewhat insufficient.'

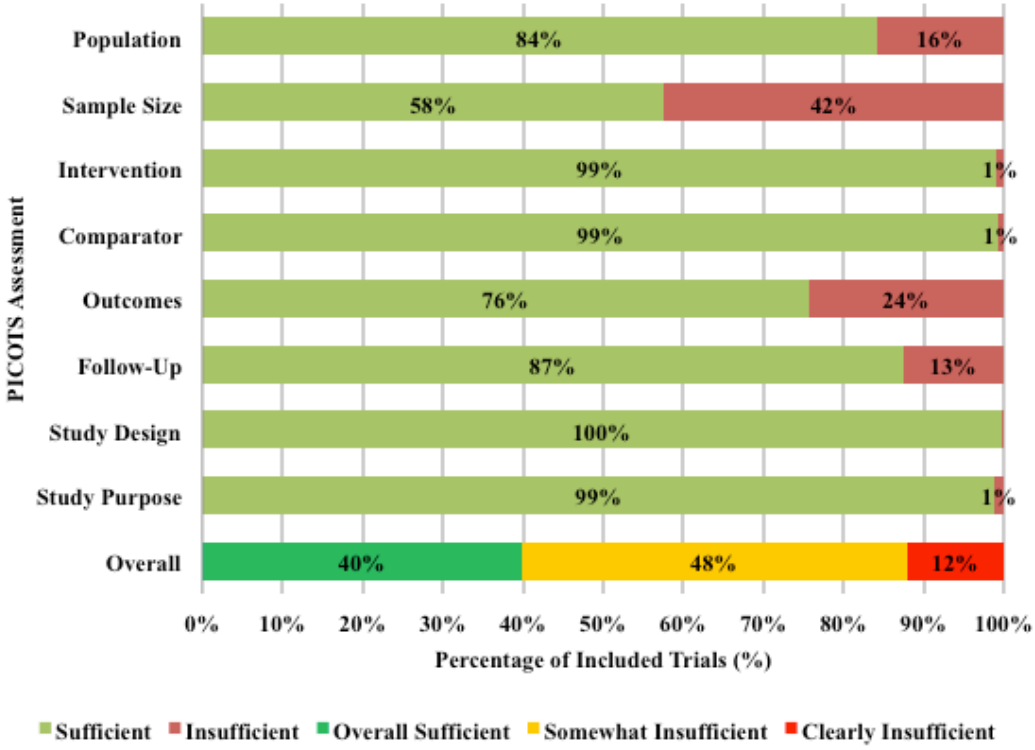


Figure 5.2 PICOTS components for all 611 included trials

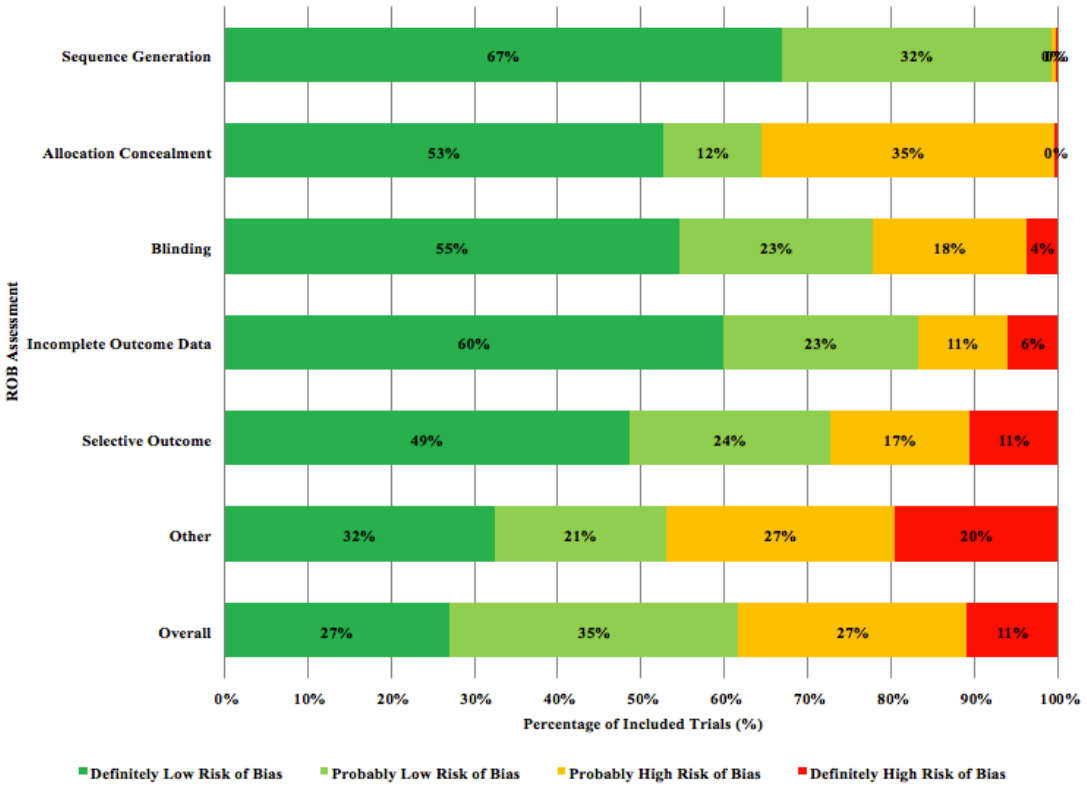


Figure 5.3 ROB components for all 611 included trials

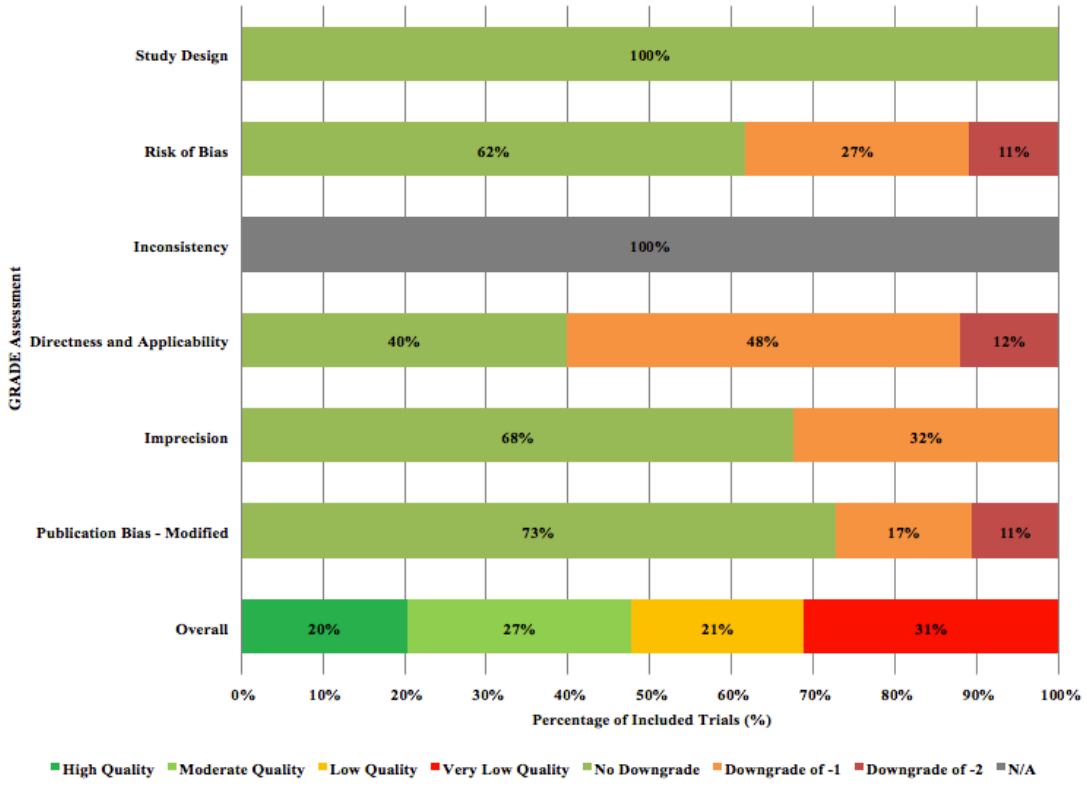


Figure 5.4 GRADE components for all 611 included trials

5.2 UNIVARIABLE AND OVERALL LOGISTIC REGRESSIONS

The relationship with reversal for each potential predictor was tested twice: individually and controlling for all others. In the univariable analyses – the results of which are presented in Table 5.2 – five of the 20 potential predictors were statistically significant at an alpha level of 0.05, and 11 had p-values less than 0.5, suggesting a relationship that did not reach significance. Imputing the missing data for four of the variables that are to be included in the overall model does not change the magnitude, or the significance, of their relationship with reversal. The beta-coefficients for each predictor in each of the regression analyses are presented in APPENDIX H. These are used to calculate the odds ratios of relevant unit differences for continuous covariates.

Table 5.5 Univariable analyses of potential predictors on "reversal vs. reaffirmation"

Covariate	(# trials / 611)	OR	95% CI	P-value
Percent participants lost to follow up (+10%)	(598)	0.83	0.69 to 0.99	0.036
Percent participants lost to follow up (imputed) (+10%)	(611)	0.83	0.69 to 0.99	0.036
Duration of follow up in weeks (+52)	(600)	0.99	0.94 to 1.04	0.592
Duration of follow up in weeks (+52) (imputed)	(611)	0.99	0.94 to 1.04	0.592
P-value (+0.10)	(535)	1.18	1.11 to 1.26	< 0.001
P-value (imputed) (+0.10)	(611)	1.19	1.12 to 1.26	< 0.001
Sample size (+100)	(611)	1.00	1.00 to 1.00	0.771
Total number of events (+50)	(473)	1.00	1.00 to 1.00	0.252
Fragility Index (+5)	(367)	1.01	0.93 to 1.09	0.866
Sufficient Adequacy of Power	(611)	1.04	0.74 to 1.46	0.826
Standardized effect size (+1)	(523)	0.89	0.83 to 0.96	0.002
Standardized effect size (imputed) (+1)	(611)	0.89	0.83 to 0.96	0.002
Year of publication (+5)	(611)	1.06	0.89 to 1.25	0.521
Years between trial start and trial registration (+5)	(412)	0.81	0.55 to 1.19	0.279
Years between trial completion and publication (+5)	(343)	1.07	0.57 to 2.02	0.835
Protocol registered	(611)	0.96	0.58 to 1.59	0.874
Abstract conclusion based on primary outcome	(611)	1.02	0.65 to 1.59	0.946
Abstract conclusion based on subgroup analyses	(611)	0.50	0.28 to 0.89	0.019
Abstract conclusion based on secondary outcome	(611)	0.85	0.62 to 1.18	0.342
Conflicts of interest	(611)			0.812
Non-industry vs. Industry		1.10	0.80 to 1.52	
None-reported vs. Industry		1.26	0.39 to 4.06	
Type of outcome	(611)			0.111
Hard vs. Surrogate		1.49	0.99 to 2.23	
Composite vs. Surrogate		1.51	0.98 to 2.34	

Overall PICOTS	(611)			0.900
Sufficient vs. Clearly insufficient		1.03	0.61 to 1.74	
Somewhat insufficient vs. Clearly insufficient		1.10	0.66 to 1.84	
Overall ROB	(611)			0.019
Definitely low ROB vs. Definitely high ROB		2.38	1.33 to 4.24	
Probably low ROB vs. Definitely high ROB		1.62	0.93 to 2.82	
Probably high ROB vs. Definitely high ROB		1.44	0.82 to 2.56	
Overall GRADE	(611)			0.477
High vs. Very low		1.27	0.81 to 2.00	
Moderate vs. Very low		1.38	0.91 to 2.10	
Low vs. Very low		1.22	0.78 to 1.91	

The covariates that we investigated were a mix of continuous, binary, and categorical. Unmodified regression results for continuous covariates are the effect of a 1-unit increase, which is not meaningful for some covariates. The effect of continuous predictors is more appropriately presented as that for a relevant unit increase. The calculations for these specific unit-difference odds ratios and their respective confidence intervals can be found in APPENDIX H.

There were two continuous covariates for which the effect was so small that a relevant unit increase failed to show an effect. For every additional 50 events in a trial, the odds of reversal, on average, do not change (OR = 1.00, 95% CI: 1.00 to 1.00 [p = 0.252]). Nor do the odds of reversal change with an additional 100 subjects (OR = 1.00, 95% CI: 1.00 to 1.00 [p = 0.771]).

Five of the potential predictors were found to be significantly associated with the outcome at an alpha of 0.05: overall risk of bias, p-value, proportion lost to follow up, standardized effect size, and abstract conclusions based on subgroup hypotheses. As the overall Risk of Bias decreases, the odds of reversal increase with each lower designation: on average, trials with an overall ROB of ‘definitely low,’ ‘probably low,’ or ‘probably high,’ had odds of reversal that were respectively 2.38 [95% CI: 1.33 to 4.24], 1.62 [95% CI: 0.93 to 2.82], and 1.44 [95% CI: 0.82 to 2.56] times the odds of trials with an overall

ROB that is 'definitely high' (p = 0.019). For every increase of 0.1 in the p-value for a trial's primary outcome, the odds of reversal increased by 19% (OR = 1.19, 95% CI: 1.12 to 1.26 [p < 0.001]). On average, an increase in the percentage of participants lost to follow up of 10% decreases the odds of reversal by 17% (OR = 0.83, 95% CI: 0.69 to 0.99 [p = 0.036]). The odds of reversal are on average 11% less (OR = 0.89, 95% CI: 0.83 to 0.96 [p = 0.002]) for every single unit increase in the standardized effect size for a trial's primary outcome. And trials with abstract conclusions based on subgroup analyses, had odds of reversal that were on average 50% less (OR = 0.50, 95% CI: 0.28 to 0.89 [p = 0.019]) than the odds of reversal compared with trials for which the abstract conclusion was not based on subgroup analyses.

Four potential predictors had p-values that were less than 0.5 suggesting a potential relationship that did not reach significance: 'years between trial start and registration,' type of primary outcome, overall GRADE, and abstract conclusion based on secondary outcomes. We describe the relationships of these covariates with reversal as "associations," based on the direction of their Odds Ratios. A 5-year increase in 'years between trial start and registration' was associated with an average decrease the odds of reversal of 19% (OR = 0.81, 95% CI: 0.55 to 1.19 [p = 0.279]). The type of primary outcome used for a comparison (p = 0.181) was associated with reversal. On average, when compared with trials using surrogate outcomes for their primary comparison, the odds of reversal were 49% (OR = 1.49, 95% CI: 0.95 to 2.34) higher for trials using hard outcomes and 47% (OR = 1.47, 95% CI: 0.91 to 2.38) higher for trials using composite outcomes. The overall GRADE quality of evidence (p = 0.477) may be associated with the outcome, as the odds of reversal, when compared trials of 'very low quality,' were

27% (OR = 1.27, 95% CI: 0.81 to 2.00) higher for 'high quality' trials, 38% (OR = 1.38, 95% CI: 0.91 to 2.10) higher for 'moderate quality' trials, and 22% (OR = 1.22, 95% CI: 0.78 to 1.91) higher for 'low quality trials.' Trials with an abstract conclusion based on secondary outcomes was, on average, associated with odds of reversal that were 0.85 (95% CI: 0.62 to 1.18 [p = 0.342]) times those of trials with abstract conclusions that were not based on secondary outcomes.

When assessed on their own, six of the potential predictors did not appear to have an association with reversal of evidence, having p-values greater than 0.50: 'years between trial end and publication,' duration of follow up, Fragility Index, Adequacy of Power, year of publication, abstract conclusion based on primary outcome, sources of potential conflicts of interest, overall PICOTS, and trial registration. We describe the relationships of these covariates with reversal as "trends," based on the direction of their Odds Ratios. On average, a 5-year increase in 'years between trial end and publication' trended towards increasing the odds of reversal by 7% (OR = 1.07, 95% CI: 0.57 to 2.02 [p = 0.835]). Every additional 52 weeks of follow up, trended to an average decrease in the odds of reversal by 1% (OR = 0.99, 95% CI: 0.94 to 1.04 [p = 0.592]). And an additional 5-unit difference in Fragility Index trended to increase the odds of reversal by 1% on average (OR = 1.01, 95% CI: 0.93 to 1.09 [p = 0.866]). On average, the trend of trials for which the Adequacy of Power was sufficient had odds of reversal that were 4% (OR = 1.04, 95% CI: 0.74 to 1.46 [p = 0.826]) higher than trials for which the Adequacy of Power was insufficient. As compared with trials published any year between 2000 and 2016, trials published an additional 5-years later trended towards reversal, with odds that were on average 6% greater (OR = 1.06, 95% CI: 0.89 to 1.25 [p = 0.521]) times greater.

For trials that based their abstract conclusion on primary outcomes, the trend was to increase the odds of reversal by 2% (OR = 1.02, 95% CI: 0.65 to 1.59 [p = 0.946]) compared with trials that did not. The sources of potential conflicts of interest (p = 0.812) trended towards influencing reversibility as the odds of reversal for trials with non-industry funding or no-conflicts reported were respectively 1.10 (95% CI: 0.80 to 1.52) and 1.26 (95% CI: 0.39 to 4.06) times that of trials with reported industry conflicts. The overall PICOTS (p = 0.900) had a similar trend as the odds of reversal were 1.03 (95% CI: 0.61 to 1.74) times higher for trials having 'sufficient' PICOTS and 1.10 (95% CI: 0.66 to 1.84) times higher for trials having 'somewhat insufficient' PICOTS, compared to trials with PICOTS designated 'clearly insufficient.' And lastly, trial registration, on average, trended away from reversal, with registered trials having odds of reversal that were 4% (OR = 0.96, 95% CI: 0.58 to 1.59 [p = 0.874]) less than trials that were not.

Before conducting the overall logistic regression, the correlation of the potential predictors with each other was checked (APPENDIX H: Table 7). As there were no highly correlated covariates – the greatest magnitude of correlation was -0.27 between 'Standardized effect size' and 'P-value' – all pre-specified predictors were included in the model. Out of the 15 potential predictors, two were significant at an alpha level of 0.05 and six had p-values less than 0.50, suggesting a potential relationship with reversal after controlling for all other predictors. All regression beta-coefficients for univariable, overall, and backwards-stepwise logistic analyses can be found in APPENDIX H: Table 8. This provides a direct comparison of the changes in covariate relationships with reversal across all analyses.

Table 5.6 Overall multivariable logistic regression (611 trials)

Covariate	OR	95% CI	P-value
Percent participants lost to follow up (imputed) (+10%)	0.91	0.75 to 1.09	0.296
Duration of follow up in weeks (imputed) (+52)	0.99	0.93 to 1.04	0.620
P-value (imputed) (+0.10)	1.16	1.09 to 1.24	< 0.001
Sample size (+100)	1.00	1.00 to 1.00	0.803
Standardized effect size (imputed) (+1)	0.93	0.86 to 1.00	0.049
Year of publication (+5)	1.04	0.85 to 1.28	0.716
Protocol registered	0.86	0.47 to 1.56	0.616
Abstract conclusion based on primary outcome	1.14	0.70 to 1.85	0.601
Abstract conclusion based on subgroup analyses	0.55	0.30 to 1.02	0.058
Abstract conclusion based on secondary outcome	0.93	0.66 to 1.32	0.696
Conflicts of interest			0.758
Non-industry vs. Industry	0.87	0.59 to 1.27	
None-reported vs. Industry	0.98	0.28 to 3.42	
Type of outcome			0.196
Hard vs. Surrogate	1.47	0.94 to 2.30	
Composite vs. Surrogate	1.45	0.90 to 2.35	
Overall PICOTS			0.640
Sufficient vs. Clearly insufficient	0.70	0.33 to 1.48	
Somewhat insufficient vs. Clearly insufficient	0.79	0.43 to 1.44	
Overall ROB			0.206
Definitely low ROB vs. Definitely high ROB	2.38	1.00 to 5.64	
Probably low ROB vs. Definitely high ROB	1.62	0.75 to 3.49	
Probably high ROB vs. Definitely high ROB	1.35	0.72 to 2.52	
Overall GRADE			0.951
High vs. Very low	0.85	0.34 to 2.12	
Moderate vs. Very low	1.00	0.49 to 2.04	
Low vs. Very low	1.03	0.57 to 1.86	

Controlling for all other covariates in the overall model, eight covariates retained the same relationship as when they were assessed on their own in that their association with prediction or protection of reversal had similar magnitude and significance. As compared with trials published any year between 2000 and 2016, trials published five years later trended to increasing odds of reversal by 4% on average (OR = 1.04, 95% CI: 0.85 to 1.28 [p = 0.716]). As the p-value for the trial's primary outcome increases by 0.1, the odds of reversal increase by 16% on average (OR = 1.16, 95% CI: 1.09 to 1.24 [p < 0.001]). A 1-unit increase in standardized effect size for a trial's primary comparison decreases the odds of reversal by 7% on average (OR = 0.93, 95% CI: 0.86 to 1.00 [p = 0.049]). Trials for which the abstract conclusion is based on subgroup analyses are

associated with an odds of reversal that are, on average, 45% lower than trials that do not (OR = 0.55, 95% CI: 0.30 to 1.02 [p = 0.058]). The type of outcome used for a trial's primary comparison (p = 0.196) may be associated with the outcome as the odds of reversal for trials with hard and composite outcomes were respectively 47% (OR = 1.47, 95% CI: 0.94 to 2.30) and 45% (OR = 1.45, 95% CI: 0.90 to 2.35) greater than for trials using a surrogate outcome. Controlling for other covariates, an additional 52 weeks of follow up trended towards decreasing the odds of reversal by 1% on average (OR = 0.99, 95% CI: 0.94 to 1.04 [p = 0.620]). The trend among trials that had an abstract conclusion based on the primary outcome compared to those that did not, was an average increase in odds of reversal of 14% (OR = 1.14, 95% CI: 0.70 to 1.85 [p = 0.601]). Sample size retained a lack of association as an additional 100 subjects neither increased nor decreased the odds of reversal (OR = 1.00, 95% CI: 1.00 to 1.00 [p = 0.803]).

After adjusting for other covariates, five had relationships in the same direction as they did on their own, but with different levels of significance. On average, trials with an overall ROB of 'definitely low,' 'probably low,' or 'probably high,' had odds of reversal that were respectively 2.38 [95% CI: 1.00 to 5.64], 1.62 [95% CI: 0.75 to 3.49], and 1.35 [95% CI: 0.72 to 2.52] times those of trials with an overall ROB that is 'definitely high' (p = 0.206). An additional 10% of participants lost to follow up was associated with a decrease in odds of reversal by 9% on average (OR = 0.91, 95% CI: 0.75 to 1.09 [p = 0.296]). Trials with registration or protocols, on average, trended towards odds of reversal that were 24% less than those that did not (OR = 0.86, 95% CI: 0.47 to 1.56 [p = 0.616]). Trials with abstract conclusions based on secondary outcomes, on average, trended towards odds of reversal that were 7% less than trials for which the abstract

conclusions were not based on secondary outcomes (OR = 0.93, 95% CI: 0.66 to 1.32 [p = 0.696]). On average, trials with an overall GRADE quality of evidence rating of 'high,' 'moderate,' or 'low' quality trended to having odds of reversal that were respectively 0.85 (95% CI: 0.34 to 2.12), 1.00 (95% CI: 0.49 to 2.04), and 1.03 (95% CI: 0.57 to 1.86) times those of trials with overall GRADE ratings of 'very low' (p = 0.951).

When all potential predictors were included in the overall model, two covariates changed their apparent relationship with reversal. As compared with trials reporting industry conflicts of interest, trials that reported 'non-industry' or 'no conflicts' trended towards odds of reversal that were on average lower by 13% (OR = 0.87, 95% CI: 0.59 to 1.27) and 2% (OR = 0.98, 95% CI: 0.28 to 3.42) respectively (p = 0.758). Furthermore, after controlling for all other covariates, the overall PICOTS trended towards decreasing the odds of reversal: compared with trials that had an overall PICOTS designation of 'clearly insufficient,' trials with designations of 'sufficient' and 'somewhat insufficient' were respectively lower by 30% (OR = 0.70, 95% CI: 0.33 to 1.48) and 21% (OR = 0.79, 95% CI: 0.43 to 1.44) (p = 0.640).

Testing the overall model using the Pearson and Hosmer-Lemeshow Goodness-of-Fit tests produces p-values of 0.210 and 0.824 respectively, suggesting that the model adequately describes the database. The overall model has 22 degrees of freedom (15 covariates and 611 trials) and consequently 15 cases (i.e. reversals) per degree of freedom.

5.3 BACKWARDS STEP-WISE MODEL

The overall logistic model was fit using all potential predictors and demonstrated that some covariates may have a significant effect on whether the results of a trial

contradict (i.e. reverse) or confirm (i.e. reaffirm) previous beliefs about the tested practice, while some may have no effect. Backwards-stepwise regression is generally not recommended for model building, but as these are exploratory analyses with no prior evidence base from which to construct a model, and with covariates that may have varying effects on the outcome, the stepwise approach was deemed suitable. All covariates from the overall model were included at the start and a dropping significance level of 0.50 was set, based on Harrell's recommendation.[130]

Table 5.7 Covariates included in the final model generated by backwards-stepwise selection

Covariate	OR	95% CI	P-value
Percent participants lost to follow up (imputed) (+10%)	0.88	0.73 to 1.05	0.152
P-value (imputed) (+0.10)	1.16	1.09 to 1.24	< 0.001
Standardized effect size (imputed) (+1)	0.93	0.87 to 1.00	0.054
Abstract conclusion based on subgroup analyses	0.53	0.29 to 0.98	0.044
Overall PICOTS			0.435
Sufficient vs. Clearly insufficient	0.69	0.39 to 1.24	
Somewhat insufficient vs. Clearly insufficient	0.81	0.47 to 1.41	
Overall ROB			0.115
Definitely low ROB vs. Definitely high ROB	2.10	1.10 to 3.99	
Probably low ROB vs. Definitely high ROB	1.52	0.83 to 2.77	
Probably high ROB vs. Definitely high ROB	1.36	0.74 to 2.48	

The final model produced by the backwards-stepwise selection included six covariates, all of which had similar relationships to reversal as found in the overall model. Two of the covariates were associated with increased odds of reversal, while four decreased the odds of reversal, thereby increasing the odds of reaffirmation.

A 0.10 increase in the p-value of a trial increased the odds of reversal of 16% on average (OR = 1.16, 95% CI: 1.09 to 1.24 [$p < 0.001$]). As compared with trials that had 'definitely high' overall ROB assessments, trials that had 'definitely low,' 'probably low,' or 'probably high' overall ROB were associated with increased odds of reversal of

respectively 110% (OR = 2.10, 95% CI: 1.10 to 3.99), 52% (OR = 1.52, 95% CI: 0.83 to 2.77), and 36% (1.36, 95% CI: 0.74 to 2.48) (p = 0.115).

On average, increasing the percent of participants lost to follow up by 10% was associated with decreased odds of reversal of 12% (OR = 0.88, 95% CI: 0.73 to 1.05 [p = 0.152]). A 1-unit increase in the standardized effect size for a trial's primary comparison was on average associated with a 7% decrease in the odds of reversal (OR = 0.93, 95% CI: 0.87 to 1.00 [p = 0.054]). Trials for which the abstract conclusions were based on subgroup analyses had odds of reversal that were 47% less than trials that did not (OR = 0.53, 95% CI: 0.29 to 0.98 [p = 0.044]). And on average, compared with trials for which the overall PICOTS assessment was 'clearly insufficient,' trials that were designated 'sufficient' or 'somewhat insufficient' were associated with 31% (OR = 0.69, 95% CI: 0.39 to 1.24) and 19% (OR = 0.81, 95% CI: 0.47 to 1.41) reductions in the odds of reversal (p = 0.435).

Testing the final model produced by the backwards-stepwise regression with Pearson and Hosmer-Lemeshow Goodness-of-Fit tests produce respective p-values of 0.345 and 0.660, suggesting that the model is adequate to describe the database. The model generated by backwards-stepwise regression has 9 degrees of freedom (6 covariates and 611 trials) and consequently 36 cases (i.e. reversals) per degree of freedom.

Figure 5.5 presents the odds ratios for all covariates across all regression analyses to show their relative magnitude, direction, and significance. The first six characteristics (overall ROB, overall PICOTS, abstract conclusion based on subgroup hypotheses, standard effect size, p-value, and proportion of participants lost to follow up) are those

from the model produced by backwards-stepwise selection and consequently have odds ratios from all three logistic regressions. The next nine characteristics (overall GRADE, outcome type, conflicts of interest, abstract conclusion based on secondary outcome, abstract conclusion based on primary outcome, protocol registered, year of publication, sample size, and duration of follow up) were excluded in the backwards-stepwise selection, but were included in the overall model and consequently have odds ratios from the univariable and multivariable logistic regressions. The last five characteristics (years between trial start and registration, years between trial end and publication, total number of events, Adequacy of Power, and Fragility Index) were those assessed only in univariable analyses and consequently only have a single odds ratio. Statistical significance ($p \leq 0.05$) of the odds ratios for each characteristic is noted on the graph as follows: * = univariable analysis, ** = multivariable analysis, *** = backwards-stepwise analysis.

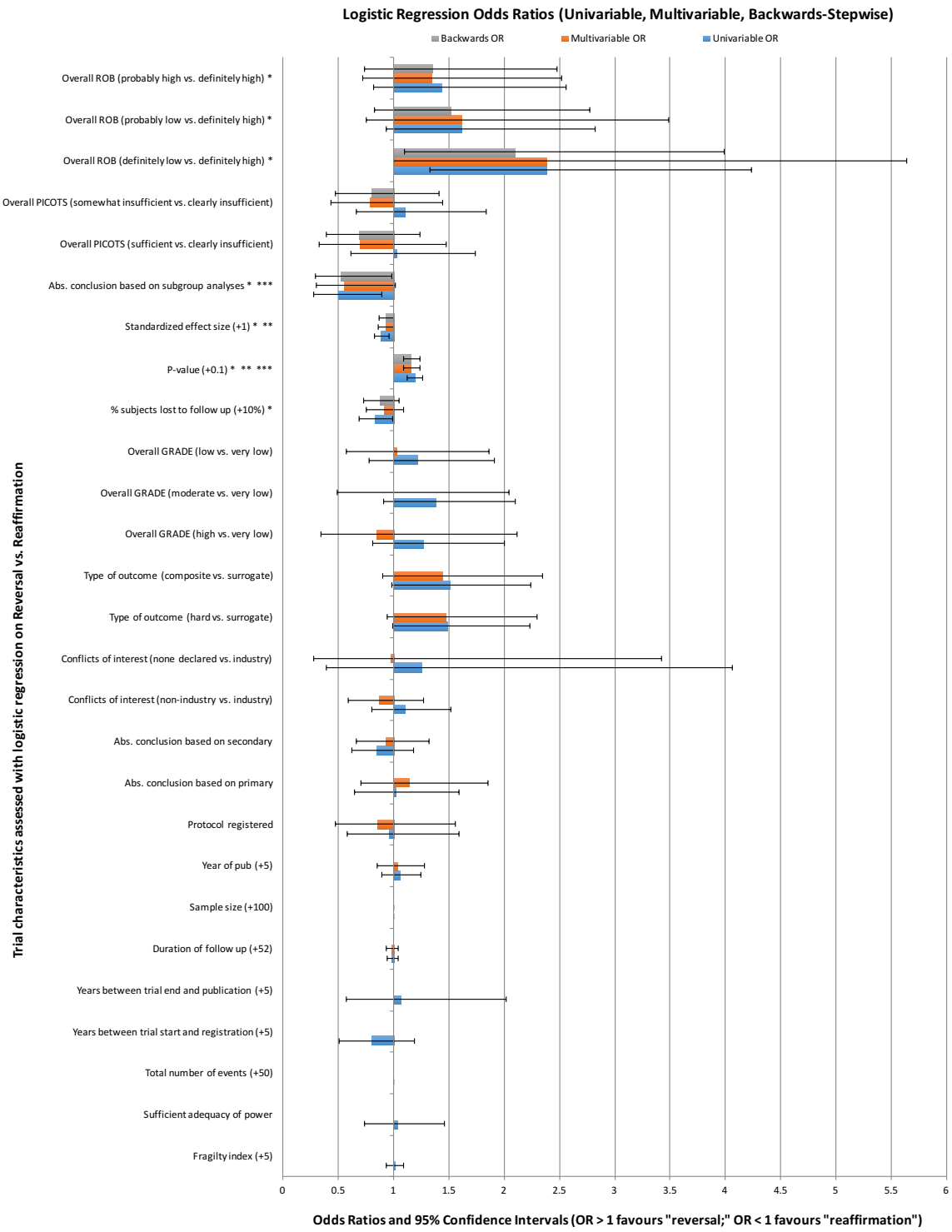


Figure 5.5 Odds Ratios of covariates across all logistic regression analyses

CHAPTER 6

A framework of reversibility

Discussion of findings and limitations

Riaz Qureshi

Chapter Summary: This chapter presents a discussion of the results of our exploratory analyses of the characteristics of reversal. A comparison is first made between our reproduction of 'A decade of reversal,' followed by a detailed discussion of the results from the logistic regressions – both expected and unexpected – and how the relationships influence the development of a framework of reversibility. The limitations of the study design and methods are presented at the end of this chapter.

CHAPTER 6

6.0 COMPARISON TO 'A DECADE OF REVERSAL'

'A decade of reversal' by Prasad *et al.* was the first major step towards understanding the phenomenon of evidence reversal. Their analyses provided rudimentary explorations of the characteristics of the phenomenon that prepared the field for more advanced study.

The trial conclusions regarding their primary outcome were based on the results, discussion, and conclusions, and incorporated statistical significance, direction of effect, and the manner in which authors described their findings. Thus, trials with statistically significant effects in favour of the intervention or in line with the hypotheses or primary question were deemed positive, trials that found statistically significant evidence contrary to their hypotheses or against their intervention were deemed negative, and trials that did not find a significant effect favouring either comparator were designated as showing no difference. We found that 42% of trials had positive conclusions, whereas 58% had conclusions that were negative or showed no difference between comparators. Prasad *et al.* found proportions of 38% and 62% respectively for trials that found the practice beneficial and trials that were inconclusive or found the intervention to be no better or worse than the comparator.[4]

With regard to the determination of trial results that contradict or confirm the use of an established standard and the consequent declaration of reversal or reaffirmation, our study showed that 54% of RCTs that tested an established medical practice represented reversal, and 46% represented reaffirmation. Prasad *et al.* concluded that 40% of studies that tested an existing practice ended in reversal, 38% in reaffirmation, and 22% were

inconclusive. This difference between our reproduction and the original likely arose because we reached a decision on all trials to force a binary outcome, whereas Prasad *et al.* deemed 22% of trials to be inconclusive.

6.1 INTERPRETATION OF REGRESSION RESULTS

In this thesis, we have performed a comprehensive analysis that is the first of its kind in the field of reversals and deepens the understanding of the phenomenon of reversal within high-quality medical literature. This thesis is the first exploration of characteristics that may be associated with reversal of established practices and as such, we have no literature upon which to base the accuracy, nor credibility of our findings. However, we know that reversals require high-quality evidence as the nature of reversal is contradicting what was previously believed, based on lower quality evidence.[1,2] Consideration of this is what led to the development of Sutton and Martin's Framework of Reversibility and the rationale for investigating our potential predictors.[45]

The Framework of Reversibility included nine characteristics of studies to consider in assessing a study's conclusions with regard to their primary comparison: PICOTS, ROB, modified GRADE, modified optimum information size, fragility index, years between trial start and registration, years between trial completion and publication, whether abstract conclusions were based on secondary outcomes, and whether abstract conclusions were based on subgroup analyses.[45] In seeking to inform the development of this framework, we assessed these characteristics and 12 others, including: proportion of participants lost to follow up, duration of follow up, p-value for primary outcome, sample size, total number of events, standardized effect size, year of publication, registration of the trial or a protocol, whether abstract conclusions were based on the primary outcome,

potential sources of conflicts of interest, and the type of outcome used for primary comparison of intervention and control groups.

The overall amount of missing data among the potential predictors was low as can be seen in Table 5.3 (Chapter 5) and the mean imputation of missing values did not affect the covariate's relationship to reversal in univariable analyses. As the imputation did not statistically change the relationships of variables with missing data, the imputed data were used in the overall model to allow the use of all observations in the database.

In describing the relationships that we found, we are aware that the majority were non-significant and the strength and validity of the relationships may be questioned as a consequence. In exploring these characteristics and their influence on finding contradictory evidence for established practices, we are aware of the low power that we have with regards to conclusions and make no claims as to declaring definitive results. Our aims were exploratory and we have been consistent in expressing the relationships that we have found as presented (i.e. significant associations, non-significant associations, or highly non-significant trends), based on the magnitude and directions of the odds ratios.

Among all of the characteristics that we investigated as potential predictors of reversal or reaffirmation, most of the relationships were as expected with a few surprising results. The expected relationships coincided with our rationale for exploring them and how we thought they might influence the likelihood of a trial reversing the established practices being tested. Some relationships were unexpected by having no impact on reversal, having an influence in the opposite direction as expected, or changing their direction of influence after controlling for other covariates. However, in controlling for

multiple covariates, we have reduced power and expect unstable estimates around the null hypothesis, which is what we see in several categorical covariates.

6.1.1 Expected relationships

The relationships that were expected included characteristics from the methodology, results, and quality of the trials. In univariable and overall analyses, trials that used hard or composite outcomes were more likely to find contradictory results than trials using surrogate outcomes. Although non-significant, the trend conforms to the expectation that using outcomes that are non-subjective is important when seeking the true effect of an intervention, as opposed to outcomes that may confirm a pathophysiological pathway, but fail to influence an aspect of health that is tangible to the patient.

We did not expect the year of publication to have an influence on reversal as there was no association found in 'A decade of reversal.' During a trial's conduct, greater proportions of participants lost to follow up lowered the odds of the trial finding a contradictory result (significantly when assessed alone and non-significantly in multivariable analyses), which is directly related to the study quality and confidence that is held in the findings. When trials have a high degree of loss to follow up, it can be difficult to differentiate between a true effect and one that is an artefact of the data that remains. Similarly, having a high value for Fragility Index and sufficient Adequacy of Power – respectively symbolizing results that are non-fragile and appropriately powered to make conclusions – are other markers for confidence in a trial's results and both characteristics trended towards increasing the likelihood of reversal. However, the relationships of Fragility Index and Adequacy of Power were only assessed in univariable analyses: Fragility Index due to a high degree of missing data and inability to impute the

missing data, and the Adequacy of Power because it was an experimental indicator that we developed. Further, as they were both highly non-significant with p-values of 0.866 and 0.826 for FI and sufficient AP respectively, we cannot draw conclusions as to their effect on a trial's reversibility.

The results of the primary comparison made in a trial also directly influenced the odds of reversal. We found the greater the p-value and the smaller the standardized effect size for a difference between comparators, the greater the odds that a trial would contradict what was previously believed about the established practice being tested. Both of these relationships were significant in all analyses and were expected, as a common trend of many reversals is high quality trials failing to replicate the results of lower-quality studies that may have found large and significant effects. These relationships with reversal have also been explored by Ioannidis as The Proteus Phenomenon: describing when early extreme results are later contradicted when attempts to replicate findings are made.[54,56] It is an established publication bias that extreme results are more likely to be published than non-significant or modest effects, which may consequently lead to early studies of practices presenting findings that are disproportionate or exaggerated from the true effect.[54]

Interestingly, in our exploration of descriptive statistics we classified the reasons for reversal and while the most common reasons fit the trend shown in the data (i.e. finding a practice 'not effective if thought to be' (46%) or 'less effective if thought beneficial/superior' (19%)), there were two others: the finding of practices to be 'harmful if thought beneficial' (19%) or 'beneficial if thought harmful/not-effective/inferior (16%).' This last reason for reversal is particularly interesting as it demonstrates the

complexity of the phenomenon and necessitates a change in the way people think of reversal. Evidence reversal does not only occur when new evidence shows a currently used practice does not work; rather, it occurs when the current belief about a practice is contradicted. Thus, a practice that is not recommended, or is recommended against, may be reversed if it is found to have a positive effect when tested, leading to its recommendation.

Many of the characteristics that we assessed were related to the quality of trials and most had the expected effect on reversal. A greater number of years between a trial's start and registration was associated with increased odds of reaffirmation. This direction of this relationship was expected as an indicator of publication biases since trials that are higher quality would be expected to have a shorter duration, with negative values indicating pre-registration and positive values indicating retroactive registration. The various sources of abstract conclusions were consistent as checks of reporting bias as trials that reported their primary outcome in the abstract trended towards being reversals, and trials for which the reported conclusions were based on subgroup analyses or secondary outcomes were associated with increased odds of reaffirmation. The size of these associations became stronger when controlling for all covariates in the overall analyses.

The relationship of the overall Risk of Bias with reversal was expected as each increasing quality level (i.e. decreasing risk of bias) had a greater effect on the odds of reversal when compared to trials with the greatest risk of bias. This monotonic relationship is significant in two regards. First, when assessed on its own, the overall categorical predictor was significantly associated with reversal. Although this statistical

significance disappeared when controlling for other covariates, the strength of all categorical associations was expected to decrease in multivariable analyses because of the reduced number of cases in each category upon which to base an estimate of effect. Second, the monotonic relationship remained throughout all analyses and, despite a loss of statistical significance, retained a similar magnitude of effect. Due to the statistical non-significance of this covariate in the overall analyses, we did not conduct a multivariable regression on the individual components of the measure. However, the apparent relationship indicates that investigating the elements of this measure in future study would be warranted and may provide insight into future applications of this assessment.

6.1.2 Unexpected relationships

While most of the relationships between trial characteristics and reversal were expected, there were some characteristics which had unexpected relationships in that they were associated with the alternative outcome than we rationalized, or their association changed direction after controlling for the effects of other characteristics.

The association of trial registration and/or use of a protocol with a greater likelihood of reaffirmation and a greater number of years between trial completion and publication with reversal were both unexpected as we believed registration and fewer years between end and publication to be integral to low risk of reporting bias. Furthermore, the relationships of other characteristics that were related to reporting and publication biases (i.e. duration between start and registration, and the sources of abstract conclusions) were as expected. The effect of registration that we found could be due to the fact that the majority of trials published before 2006 were not registered because it was not yet a standard requirement of clinical trials. Consequently, the proportion of

reversals and reaffirmations that would have been registered may be skewed and the true effect obfuscated. However, by including the year of publication in the overall analyses, the effect of changing requirements over time should have been accounted for, yet in the overall analyses the strength of the relationship grows (albeit remaining a non-significant trend) instead of diminishing or reversing. Similarly, the duration between trial end and publication was only available for some trials that were registered and consequently had a high proportion of missing data (44%).

We believed that a greater duration of follow up would be associated with a greater likelihood of reversal as most interventions require a long period of follow up to determine their true effect. However, we found that the greater the duration of follow up, the odds trended towards reaffirmation – though this effect was small and non-significant: an increase in follow up of 52 weeks increased the odds of reaffirmation by 1% in both the univariable and overall analyses. Also unexpected, for the same reasons, were the effects of sample size and total number of events. For each additional 100 subjects or 50 events in a trial, there were no trends in changing in the odds of reversal (OR = 1.00). Both of these characteristics are classically portrayed as being paramount to quality trials as they decrease the variability in average outcomes and present more accurate portrayals of interventions effects. It is possible that no association was found for these characteristics due to the population from which they came. In looking at the difference in mean duration of follow up and sample size among reversals and reaffirmations (APPENDIX H: Table 6), reaffirmations are 7.28 weeks longer and reversals are 273 subjects larger, which may not be a large enough to establish an effect on the outcome. It is possible that this similarity could derive from publication in a high-

quality medical journal, which may accept trials that are of similar size and conduct, thereby rendering some characteristics uniform across all trials, regardless of results and conclusions.

Among the unexpected results were three potential predictors that exhibited the expected relationships when analyzed on their own (increasing the odds of reversal), but changed to increase the odds of reaffirmation when controlling for other covariates in the overall analyses. These included: potential conflicts of interest, the overall PICOTS assessment, and the overall GRADE quality of evidence.

While none of the above characteristics were significantly associated with reversal or reaffirmation in any of the analyses, they all trended towards increasing the odds of reversal on their own, and increasing the odds of reaffirmation when adjusting for all covariates. While the change in direction of these relationships in the multivariable analyses was interesting, it was unsurprising as they are all categorical covariates and we expected low power and unstable estimates around the null for these covariates when controlling for other predictors, as the numbers of trials in each category upon which to base an estimate are reduced. We expected the declaration of potential conflicts of interest as 'industry' would lead to greater odds of reaffirmation compared with 'non-industry' or 'none to declare/reported,' which is the trend we saw when analyzed as a single predictor. We did not expect a monotonic relationship in this covariate as the categories were nominal, not ordinal. It is likely that the change in effect was due to the instability of estimates around the mean, but it is also possible that there may be other characteristics that are influenced by the presence of conflicts of interest (e.g. sources of

abstract conclusions, type of outcome used, standardized effect size, or overall PICOTS or ROB) that consequently confound the effect in multivariable analyses.

The overall PICOTS and GRADE assessments are multidimensional summary scores for trials that each account for several elements of a trial's design and conduct. Consequently, while higher quality levels of both, on their own, are associated with increasing the odds of reversal, it is possible that when controlling for all other covariates – some of which may influence the components that make up the summary scores – the expected effect is lost. Furthermore, neither summary score had a consistent monotonic relationship with reversal as found with the Risk of Bias assessment. In univariable analyses, the highest quality PICOTS and GRADE assessment both trended towards increasing the odds of reversal to a lesser degree than assessments of moderate quality. While the non-significance of these trends is important in knowing the limitations for how we draw conclusions regarding the effects of covariates, the fact that we did not find a clear effect for GRADE in even the univariable analysis is interesting. Since we modified GRADE for application with a single trial – from its validated use in assessing aggregate evidence – its relationship to contradictory results may be less appropriate than that of a measure designed for a single trial (such as ROB). However, as GRADE is considered to be the gold-standard for assessment of evidence, the lack of an apparent relationship with reversal demonstrates a need for further exploration on more appropriate and larger datasets.

6.2 UPDATING THE FRAMEWORK OF REVERSIBILITY

The framework developed by Sutton and Martin, based on assumptions about the nature of trials that lead to reversal, was comprised of eight components: PICOTS, ROB,

and modified GRADE assessments, modified optimum information size, fragility index, duration from trial start to registration and from completion to publication, and abstract conclusions.[45]

In extracting these characteristics for testing, the optimum information size was amended to Adequacy of Power. The difference being that optimum information size is a concept from meta-analyses that was difficult to apply to individual trials – in essence: a means of assessing whether or not a meta-analytic database had sufficient power, based on the assumption that the overall sample size was equivalent to a trial of the same size – but the Adequacy of Power was more applicable to single trials and derived more simply from whether the trial met its pre-specified sample size and whether the confidence limits of the effect met the trial's pre-specified delta.

From our univariable analyses, the fragility index, sufficient Adequacy of Power, and duration from trial completion to publication do not appear to influence the likelihood of reversal in a meaningful way, so we can remove these from the framework. While a measure of the fragility of a trial's findings is an interesting concept, and deserves further study on its own, it may not contribute meaningfully to our framework of reversibility given its overlap with other concepts inherent to the framework. Additionally, providing context for the FI (e.g. as a percent of loss to follow up or the sample size), could increase the meaningfulness of this measure.

The overall GRADE assessment was highly non-significant and did not demonstrate a coherent relationship with reversal across the analyses. For this reason, we also remove the modified GRADE assessment from our framework.

Both the overall ROB and PICOTS assessments were demonstrably associated with reversal and reaffirmation respectively, such that they were included in the final model created by the backwards-stepwise regression. It is possible that these two characteristics – themselves being two of the components that contributed to the final GRADE score – are sufficient measures of study quality and bias and that GRADE is not additionally needed in considering the likelihood of reversibility. As such, both of these summary scores remain in our framework.

The years from trial start to registration was only assessed in univariable analyses but was found to have a potential association with reaffirmation as the p-value was 0.279 and the corresponding OR for reversal was 0.81 for an increase of five-years. While non-significant, the trend is expected and Harrell suggests that a model built on theory is more purposeful than one based solely on the data and that the exclusion of all non-significant predictors is often inappropriate.[130] As such, this covariate remains in our framework.

Similar justification applies to the retention of the sources of abstract conclusions from the original framework to the updated. Although only one source (subgroup analyses) had a significant relationship with whether or not the trial found contradictory evidence, the other two (primary or secondary outcomes) trended towards the outcome that we expected when controlling for other covariates to a greater degree than when they were assessed on their own, and together they provide a complete picture of the source of a trial's abstract conclusion.

After conducting our exploratory analyses of trial characteristics, we found several covariates that are strongly associated with reversal or reaffirmation of established practices and we are consequently adding them to the framework. The

relationships of the p-value and standardized effect size for a trial's primary comparison were both significant in univariable and multivariable analyses. The larger the effect found by a trial, and the more significant the result, the more likely that trial was to confirm an established practice. The same association with reaffirmation was found for increased proportions of subjects lost to follow up. Although it was only significantly associated when it was assessed on its own, it was one of the final covariates included in the model created by the backwards stepwise regression. The final characteristic that we explored and are adding to the updated framework is the type of outcome used for a trial's primary comparison. This covariate was not statistically significant in either the univariable or the multivariable regressions, but the association with reversal remained almost unchanged both in magnitude and significance between its baseline effect and after controlling for other covariates. Furthermore, this is one of the only predictors for which an evidence base existed to support its association with evidence reversal as it is known that surrogate outcomes do not necessarily correlate with patient important outcomes. Because of this, trials that test established practices with hard and definitive outcomes may contradict what was believed about practices that were prematurely adopted based on results from studies using surrogate outcomes.

Our updated, proposed framework of reversibility is presented in Table 6.1 and includes eight components that are supported by our results as being associated with the outcome of contradictory results among randomized controlled trials that test established medical practices. These components can separately be placed into five of the domains of a randomized controlled trial: design, conduct, results, quality, and reporting.

Table 6.1 Updated proposed framework of reversibility

Component	Purpose in the Framework
Overall PICOTS	Multidimensional summary score of the appropriateness of a trial's question and design
Overall ROB	Multidimensional summary score of a trial's quality
Years between trial start and registration	Measure of reporting bias as trials should be registered before they begin (i.e. have negative values)
Sources of abstract conclusions	Measure of reporting bias as trials should base abstract conclusions on primary outcome, not secondary outcome or subgroup hypotheses
P-value	Measure of significance of a trial's findings for primary comparison
Standardized effect size	Measure of the magnitude of effect for primary comparison
Proportion of subjects lost to follow up	Indicator for the confidence that can be held in a trial's results
Type of outcome	Indication of the use of an appropriate outcome for finding the clinical effect of intervention

6.3 STRENGTHS AND LIMITATIONS

There are many strengths in our study design and conduct, the most prominent being that this is the first quantitative assessment of study design elements and their relation to evidence reversal in this newly emerging field of meta-research. This study is also the largest and most comprehensive examination of the phenomenon to date. The similarity of results that we found in the reproduction of 'A decade of reversal' is another strength that lends validity to our methods. The thoroughness and care that was taken in ensuring accurate and consistent data extraction is a major strength of our study. Similarly, our high degree of transparency is a major strength as we have provided all methods used for data extraction with the intention of increasing the reproducibility of our findings for future researchers. However, there are still limitations in the design and conduct of our study. The limitations that could be addressed were, to the best of our ability, but there were still some that could not be addressed because of time constraints, a lack of resources, or necessitated assumptions based on practicality and feasibility.

6.3.1 Limitations in the creation of the database of reversals

- A critical limitation is the use of a single journal (NEJM) as a source of trials and potential reversals. An ideal search would have collected RCTs from several different journals or databases. However, given the aims of this study and limited resources for conducting the study, a single-journal was deemed most feasible and appropriate. It does however limit the generalizability of our findings to other medical literature where the majority of studies are published, and upon which many health-care decisions are made.
- The lack of time and resources led to several other limitations including the first two levels of screening and the data extraction not being done in duplicate. It also contributed to a difficulty in reaching decisions with regard to the established use of practices, as the reviewers do not have clinical experience and did not have the time to conduct literature searches to verify the existence or novelty of every practice. While a decision was reached for each trial's intervention being new or existing, and with regards to the outcome being reversal or reaffirmation of evidence, the decision was not always clear. Despite the discussion of discrepancies and reaching agreement as to what was believed to be the correct designation, some readers may disagree with how articles were categorized. We tried to be as objective as possible in our determination and to guide the decisions by what was provided by the publication authors. Thus, despite this limitation, we are confident in our results and feel that a few disagreements from other meta-researchers would not likely change our conclusions.

- Another limitation that is related to our objective judgment of articles is the extraction and assessment of PICOTS and ROB. We attempted to prevent potential bias and ensure uniformity across extractions in two ways: firstly, by extracting a test-set of trials in duplicate and comparing extraction to verify the similarity of results; and secondly, by having a protocol with clearly defined guidelines for how to extract all elements in the database. However, these measures are subjective and even trained assessors may apply different ratings of overall sufficiency and likelihood of risk of bias.
- This same limitation also applies to GRADE because even though the rating consistency was mediated slightly by the automatic completion of components based on other extracted characteristics, the automatic completion of its individual domains were taken from other subjective characteristics (i.e. overall PICOTS, overall ROB, ROB for selective outcome reporting, and the adequacy of power.).
- A further limitation of the GRADE assessment was that it was incomplete. We used a modified GRADE that did not include the domain of inconsistency because it would not have been feasible to examine the relevant literature for each intervention to assess inconsistency of results within that field, nor appropriate to compare the inconsistency of results across the many different types of interventions that we included in our review. The GRADE component for publication bias was also modified to be determined by the likelihood of selective outcome reporting because we did not have time to explore the relevant literature for each trial's comparators.

- A less consequential limitation of our study conduct was the use of an Excel file for our screening and the creation of our database. As a general rule for any study involving large amounts of data collection, this practice should be avoided because of the potential for transposition errors and incorrectly encoding automatic columns. However, the ease of use, availability and access to the program, and versatility for analyses across multiple types of quantitative and qualitative information made it ideal for this project.

6.3.2 Limitations of statistical analyses

- A major limitation of our analysis is the designation of reversal or reaffirmation of evidence based on the author's declaration and description of how their findings align with current beliefs about the practice (i.e. whether their findings are incongruent or congruent with practitioners' use). Ideally, cumulative meta-analyses or a similar measure of sufficiency and stability of the evidence for a practice would be used to indicate reversal or reaffirmation. As currently designated, these author's conclusions may yet again be contradicted with time and we do not know how the practices and standards have changed since the trials that we have used were published. However, populating our database with RCTs and not meta-analyses or other standards for decision-making is another consequential limitation. Prasad *et al.* examined all original studies and so we too looked at original studies, but chose to examine RCTs because they are the gold standard for investigating the effects of interventions and should provide a higher level of evidence for decision-making than other original research designs.

- Another limitation related to the outcome used in our regression is that all reversals, partial and full, were categorized the same. It would have added an extra layer of complexity that would have made the project infeasible to try and distinguish between complete reversals (e.g. findings that indicate direct harm and a recommendation for immediate cessation of use) and partial reversals (e.g. findings that a practice does not work as well as was believed, but it still has use and a recommendation for further study). Similarly, when a trial tested two established practices where there was no consensus about which is better (as was often the case), and one was found to be superior, then it was classified as a reversal because it contradicted the belief that they were equal, even though in doing so they also confirmed the use of the superior practice.
- This difficulty in classifying reversals and reaffirmations of evidence introduces a further limitation by necessitating an assumption for our logistic regressions: namely that the outcome follows a binary distribution. We reached a decision on all included trials, but Prasad *et al.* deemed 22% of studies that test established practices to be inconclusive – neither reversing, nor reaffirming a practice. While the differentiation between reversal and reaffirmation of evidence is not black and white, we consider the assumption valid for the purposes of exploring the phenomenon of reversals given that most of the decisions for reversals will never be without some uncertainty, and most of the decisions could be adequately inferred from the original studies' details.
- Within the analyses – both descriptive and regression – there are also several other assumptions that are limiting but were necessary to allow the analyses to

occur. The most prominent of these is that any error in data extraction is random and not systematically biased. It was assumed that the judgments made by data extractors (RQ and DS) were comparable with respect to PICOTS, ROB, and consequently GRADE. As explained in section 6.3.1, we attempted to mitigate differences in extraction by comparing responses across a random test-set of articles before completing the extraction and by using a pre-specified protocol (APPENDIX E) to guide the extraction.

- Another important limitation in the analysis and extraction arose from the inclusion of trials with multiple interventions including multi-arm studies and factorial designs. Within our study sample, there were 98 trials with these designs, comparing multiple practices against each other and control groups. For these trials, we extracted only a single intervention and control group for our summary of study results. This simplifying assumption was necessary for project feasibility as extraction of every pair of comparisons from the multi-arm and factorial trials would have increased the number of "trials" to unmanageable levels. The decision to include only a single comparison from each trial publication also acts to simplify the analysis as the number of "subjects" that are related is greatly diminished and the independence of our observations can be assumed.
- An assumption made in the overall logistic regression analyses included that all measures of effect were directly comparable through the use of a standardized effect size. Not all event data is directly comparable when accounting for the primary question (e.g. if a trial intended to find an intervention's effect on time-to-event, then comparing the overall event rates between groups may not be a

valid indicator of the effect in question). Despite this limitation, an assumption was made that comparability through the use of a standardized effect size – based on the Absolute Risk Difference for dichotomous outcomes, calculated using raw data, and based on the mean difference for continuous outcomes – would be appropriate to convey the relative magnitude and direction of effect, seen across all trials.

- Calculating the Fragility Index for all trials with dichotomous outcomes that compared two interventions in 1:1 allocation involved an assumption that all dichotomous event data were comparable, as it is calculated using the numbers of events and subjects in each group. This is a limitation as the applicability of the Fragility Index to hazard data and trials with long periods of follow-up has been questioned due to the tendency of both types of trials to reach similar overall event rates in groups with increasing time.
- Another assumption of the regression analyses was the interpretation of "Not Available" as missing data for covariates, and that the overall amount missing does not affect the outcome of the analyses. As a result of this interpretation, the amount missing for some covariates was exaggerated. 'N/A' was used in the database for information that could not be found within a trial (e.g. some trials did not report a p-value for their primary outcome), and it was also used to denote when a response was not possible for a particular covariate (e.g. the Fragility Index cannot be calculated for trials that have continuous outcomes, compare more than two interventions, or have allocation ratios other than 1:1).

- A major limitation with our analyses lies in our choice of backwards-stepwise model selection for our exploration of characteristics that may be associated with reversal. Stepwise model selection procedures are generally not recommended for building predictive models because they are unreliable in the presence of collinearity, lead to high-biased R^2 values, generate standard errors for the parameter estimates that are too small and consequently parameter confidence intervals that are too narrow, create parameter estimates that are biased high in absolute value, and generate p-values that are biased low.[130] Stepwise selection procedures can be appropriate for exploratory analyses when there is no prior information to guide variable selection. However, while traditional significance criteria for exclusion from a model tend to be stringent (e.g. $\alpha = 0.05$ or 0.10), an alpha of 0.5 is more reasonable in allowing for the deletion of some variables that may be irrelevant to the outcome, but the retention of most variables that may help to predict the outcome, despite insignificance.[130] Given that our purposes were exploratory and there is no literature to guide variable pre-specification in relation to reversal, we believe a backwards-stepwise model selection to be appropriate for informing the development of our framework of reversibility.
- A final limitation in our interpretation of results is the reporting of relationships on the basis of the directionality and magnitude of covariate odds ratios. We are aware that the majority of potential predictors were not statistically significantly associated with our outcome, but as these analyses were exploratory, we felt that it was best to describe and interpret the relationships that we found in terms of whether they trended towards influencing the outcome in the way that we

expected, whether or not they were significant. In doing so, we referred to covariates that were highly non-significant ($p > 0.5$) as trending relationships, covariates that were moderately non-significant ($0.5 > p > 0.05$) as associated, and covariates that were significantly associated as such. This is a limitation in that we make no claims to these covariates being definitive predictors of contradictory findings for established practices, but we present the magnitude and degree of associations as we found them and make recommendations for which relationships we believe further study or consideration in use is warranted.

CHAPTER 7

The future of reversibility research

Impact, Applications, Future Directions, and Conclusions

Riaz Qureshi

Chapter Summary: This chapter presents further discussion of the overall results of this study, particularly with regards to the impact and applications of this research and the future directions that should be explored in the field of reversals. The chapter ends by presenting an overall summary of the thesis and its conclusions.

CHAPTER 7

7.0 IMPACT AND IMPORTANCE

Our framework of reversibility consists of eight components of randomized controlled trials that have relationships with the likelihood of reversal and cover multiple domains from design to reporting. The development of a framework of characteristics that should be considered in assessing trials that contradict current standards and established practices has important implications for the field of reversals and to our knowledge has not previously been attempted. Not only is this the largest review of the phenomenon of reversal – updating the previous largest review with an additional seven years of studies – but also significantly expands analyses of associations between study variables and reversal. This is the most comprehensive, and the first quantitative, exploration of trial characteristics that may lead to evidence reversal. This framework can serve as a tool to be used by researchers and health policy-makers in guiding the decisions around adoption, de-adoption, and dis-investment of practices.

7.1 APPLICATIONS OF THE FRAMEWORK

The responsibility of identifying evidence reversals lies primarily with the researchers and developers of interventions and standards, and also with agencies that grant approval for research and implementation.[2,26,36] In knowing where the burden of proof lies, consideration must be given to the methods and tools that are currently employed for identifying and reducing the impact of evidence reversals and medical reversals. While standards of practice exist in all scientific disciplines, and new findings always require dissemination before they can be implemented, there currently exist several methods by which the effect of reversals are mitigated and which may benefit

from incorporating our framework of reversibility. These are clinical guidelines, knowledge translation, and various other tools for de-implementation of practices.

7.1.1 Clinical guidelines

Evidence Based Medicine is a common goal for decision-makers in literature, yet practice does not always follow the best available evidence. Physicians report that the proportion of their practices that are evidence based are as low as 50% of their practices.[141] This proportion is unsurprising given the overwhelming amount of literature available to physicians and the difficulty that exists in finding clinically important literature that is relevant to practice and has enough evidence to inform a decision.[142] A major systematic review of the publication of clinically important and relevant articles in primary healthcare journals suggests that clinicians would need to read an average of 13-14 articles to obtain one that is directly clinically relevant.[143]

As a consequence of this information overload, many clinicians rely on clinical guidelines in medical practice.[141] However, the quality of clinical guidelines varies significantly based on the methods and processes used to select and apply evidence to guide recommendations. As a result, the benefit of clinical guidelines is somewhat contested and they are not guaranteed to reduce the impact of medical reversals.[144] Despite their necessity for efficient and consistent practice, there remain several inherent difficulties in generation and dissemination of effective and unbiased guidelines.[144] These challenges may be especially powerful barriers against effectively mitigating the impact of evidence reversals.

The first difficulty in establishing a clinical guideline is creating an impartial team where there is no conflict of interest. A survey from 2012 found that 71% of chairs of clinical policy committees and 90.5% of co-chairs had financial conflicts of interest and

that these conflicts could have a substantial effect on the conclusions and recommendations of the guideline.[145]

The second major difficulty with guideline creation is missing data as a direct result of poor knowledge translation practices (i.e. publication bias) and proprietary rights of those who own data.[38,146] One of the outcomes of poor knowledge translation that contributes to the difficulty in establishing clinical guidelines is a general lack of confidence in conclusions of efficacy.[147] An evaluation of the quality and sufficiency of evidence for clinical practices by The Cochrane Collaboration in 2011 found that as many as 45% of Cochrane reviews conclude that there is insufficient evidence to endorse the intervention.[147] Due to insufficiency of evidence, an analysis of clinical care in Australia found that patients only received appropriate care (i.e. based on expert recommendations and clinical guidelines) between 54% and 57% of the time.[148] Increasing the amount of open data and transparency in research findings would be the first step to improving clinical guideline development and reducing unnecessary medical reversals.[38,149]

Beyond the difficulties associated with conflicts of interest and knowledge translation, clinical guidelines are also controversial because they can become out of date very quickly, they often require incredible resources to assimilate all the relevant information, there are often overlapping guidelines to consider, they are often written for a general population but must be applied to patient-specific needs, and they are often not sensitive to local needs or circumstances wherein the decision maker might otherwise have used the available evidence and their clinical expertise to devise a more appropriate therapy.[150,151]

Our application could be used as a tool in the development of clinical guidelines, in conjunction with other methods, when considering new evidence that contradicts established practices. In order to declare a reversal of evidence and recommend the de-adoption or adoption of a practice, the totality of evidence in support of it must be considered. Our framework suggests a set of characteristics of RCTs that may be associated with evidence reversal and should be considered when assessing trials for informing changes in recommendations for practice. Furthermore, as the 'evidence base' for evidence reversals matures to define predictors, future guideline developers should consider risk of reversibility of the evidence before they recommend implementation of a new intervention.

7.1.2 Improved knowledge translation

Poor knowledge translation (especially premature translation) is directly related to evidence reversal.[118] According to Prasad, "translation failure occurs when the results of preclinical, observational and/or early phase studies fail to predict the results of well done (i.e. appropriately controlled, adequately powered, and properly conducted) phase III or randomized clinical trials."[118] While knowledge translation and de-implementation go hand in hand with medical reversals, there are barriers to knowledge translation and reasons why good evidence is not readily adopted into clinical practice.[152] Barriers may include characteristics of the evidence itself, features of the practice environment (e.g. financial disincentives, organizational constraints, perception of liability, patient expectations), the prevailing opinions and social contexts for treatment (e.g. standards of practice, beliefs held by opinion leaders, out-dated medical training, advocacy groups), or knowledge and attitudes about interventions in the professional context (e.g. clinical

uncertainty, physician's sense of competence, compulsion to act, information overload).[153,154]

If knowledge translation could be improved and clinicians made more aware of which treatments had proven efficacy and which did not, it logically follows that there should be a decrease in the use of harmful or unnecessary practices. However, there are a number of reasons why physicians may continue to use treatments that are harmful or do not work, including: clinical experience, over-reliance on a surrogate outcome, natural history of the illness, strong belief in the pathophysiological model, ritual or mystique (i.e. medical tradition), a need to do something and take action, patients' expectations, or even because the correct questions about the treatment have not yet been asked.[52] The continued use of practices that should have been phased out is common as evidence suggests that up to 25% of patients receive treatments that are harmful and as many as 40% receive treatments for which the effectiveness is not known or inconclusive.[155,156]

The assumption that improved knowledge translation may decrease the premature adoption of practices may be questioned with the argument that publication bias leads to an incomplete understanding of any given research topic, with preferential publication of significant and positive findings.[39,157] This argument would infer that improving the translation of research findings would not necessarily be commensurate with a decrease in the impact of unnecessary reversals if the research that is being translated is being published with bias or 'false' due to other reasons, such as p-hacking, outcome switching, or newness of a field.

In application to knowledge translation, our framework comprises elements of high quality studies at all stages – from conception to publication – that we found to have

potential relationships with the conclusions of RCTs that test established practices and should be considered by investigators in designing trials of established practices.

7.1.3 De-implementation tools

Without knowledge translation, the development of evidence-based clinical guidelines would be impossible. These tools for EBM are primarily thought of as "positive re-enforcers" (i.e. guidance on what to do in practice), useful for the implementation of practices and standards. However, both knowledge translation and practice guidelines can also be negative and serve as tools for de-implementation (i.e. guidance on what not to do in practice).

A number of campaigns have been proposed to aid in the de-implementation of practices that have been, or are likely to be, reversed. Practices that are the target of de-implementation are often low value or not supported by the evidence. These tools exist to provide summary recommendations for the cessation or continuation of practices based on the best available evidence. They include: the United Kingdom's (UK) National Institute for Clinical Excellence (NICE) "do-not-do" lists and Database of Uncertainties about the Effects of Treatments (DUETs) [116,158]; Canadian, American, and UK "Choosing Wisely" Campaigns [70,95,106]; the British Medical Journal's (BMJ) "Too Much Medicine" [78]; and Australia's "Low-value lists."[66]

Similarly to the development of guidelines, our framework may have use in helping inform the de-implementation of practices that are harmful or of low value. It provides a set of characteristics that should be considered when assessing trials that test established practices.

7.2 A TOOLBOX FOR FUTURE REVERSAL RESEARCH

In the pursuit of reducing the burden and harms associated with unnecessary medical and evidence reversal, there are two important characteristics to consider. The first is the identification of practices or paradigms that are to be tested. There are many global initiatives that list contradicted, unproven, or new interventions and claims. These categories provide potential targets for reversal and de-implementation and promote awareness to practitioners regarding the maturity of interventions that they use in daily practice, as many have been adopted prematurely based on inadequate evidence.[44] The second characteristic is whether or not enough evidence has been accrued to confidently make a decision regarding the tested paradigm. One method that has been utilized for public health interventions, and which we propose would be appropriate for describing the evidence base for reversals, is the calculation of evidence sufficiency and stability.[159]

At their core, the characteristics of sufficiency and stability provide a means of determining the point at which an intervention has been studied enough that conducting another test no longer provides any information of added value for decision-making.[159] When interventions are studied in humans, it is ethically irresponsible to conduct research beyond this point as participants will be unnecessarily randomized to receive no benefit (or harm if the intervention is determined to be dangerous), the implementation of effective risk-reduction interventions will be delayed, and there may be unnecessary waste of health care resources.[159]

Evidence sufficiency refers to whether or not a meta-analytic database demonstrates that an intervention does or does not work to an adequate degree.[159] However, evidence sufficiency is a term that requires further exploration, especially with

respect to its relationship to tracking evidence cumulation and ultimately indicating when "enough evidence" has accrued, because it has been inadequately discussed in the literature and formal definitions still need to be derived. Some have suggested that sufficiency refers to whether or not cumulation such as through meta-analysis demonstrates that an intervention works, or does not work, with sufficient margin of difference and with sufficient precision (narrow CI) to suggest further evidence is unlikely to change this conclusion. Sufficiency may be related to the number of hypothesis tests (i.e. studies) and the power within those studies.[159,160]

Evidence stability refers to the shifts in direction over time for support of the intervention being studied.[159] Stability derives from the flow of the running estimate generated over the sequential meta-analyses: if the evidence all tends to point in one direction then the database is stable.[159,160]

These definitions represent initial attempts at defining the concept of when there is 'enough evidence', and should be further defined and tested to determine whether they can be better used to inform the concept of evidence reversibility. There are several different meta-research methods that have been proposed to describe evidence sufficiency or make use of the maturity of evidence in medical decision-making, including Value of Information Analysis, Trial Sequential Analysis, and Bayesian analysis. We propose that these methods, which we describe in APPENDIX I, be explored as tools for use in future research about reversals, particularly with regards to determining the sufficiency and stability of the evidence base for practices that should be de-adopted, or those for which the evidence is mature enough to warrant a recommendation of adoption into practice.

7.3 NEXT STEPS AND FUTURE DIRECTIONS

With this thesis, we have provided a comprehensive overview and study of a phenomenon that is rapidly gaining attention in the medical community and beyond. Through literature reviews and detailed exploratory analyses, we have presented the first quantitative examination of characteristics that may be associated with reversibility within high quality medical literature. However, this thesis is only one step of the many required before understanding the phenomenon and being able to reduce the impact of reversals on population health.

Unnecessary reversal implies several harms to the medical industry and the patients who are administered treatments including decreased trust in the medical community, the possibility of receiving unnecessary or ineffective treatments, and an increased risk of unnecessary harm.[26] The most efficient way to minimize these harms would be to reduce the impact of evidence reversal and medical reversal that occurs as a consequence of practices being prematurely adopted: before the evidence has sufficiently demonstrated the true effect. If practices were adopted only after the evidence for their use had matured, then the rate of unnecessary reversal would be lower.

This thesis has been a comprehensive analysis of the contradiction of evidence by individual trials, and builds upon the foundation of the first area of research in the field (identified in our Chapter 2 literature review): research about the phenomenon itself. There are already many initiatives worldwide that pursue the second major area of research (i.e. the practices that are low value or harmful and should be targeted for de-adoption and reversal), and our proposed toolbox for reversal includes several methods that we believe should be further explored for use in determining the sufficiency and

stability of an evidence base to declare a practice as being reversed or a confirmed standard of care. One of the next steps for reversal research should be the exploration of de-adoption and de-implementation strategies for practices that should not be in use. This will be challenging as some practices will have evidence mature enough to support or refute their use, but many will have immature evidence to adequately support their use, which will require judgement calls about whether likely benefits outweigh the risks and costs of continued use. Another major next step for research in reversal will be to explore the potential predictors of future reversal – the characteristics of original research that lead to practices being prematurely adopted – and to create a predictive model for the likelihood that a newly adopted developed practice may be reversed in the future.

7.4 CONCLUSIONS

Evidence reversal can be defined as "when new evidence that is stronger than preceding evidence arises to contradict previously established evidence." In our analysis of 17 years of original studies from the NEJM, a total of 54% of randomized controlled trials that tested established practices met the definition of evidence reversals. Within these trials, a total of 8 characteristics were associated with reversal including: overall PICOTS and ROB assessment, number of years between a trial initiation and registration, sources of abstract conclusions, p-value and standardized effect size for the primary comparison, proportion of subjects lost to follow up, and use of surrogate versus clinically-relevant outcomes. Using these characteristics, we propose an Evidence Reversal framework which may be useful for tracking and detecting evidence reversals, and for informing design of future robust RCTs that challenge established practices.

These results provide a research agenda to better inform related research for the rate and predictors of reversal and for identification of low-value practices as targets for de-adoption. Perhaps more importantly, this research may help to better inform next steps towards preventing premature adoption of new treatments, which represents a significant driver of inefficiency and waste in healthcare.

REFERENCES

1. Prasad V, Cifu A. The frequency of medical reversal. *Arch Intern Med*. 2011;171(18):1675-1676. doi:10.1001/archinternmed.2011.295.
2. Prasad V, Cifu A. Medical reversal: Why we must raise the bar before adopting new technologies. *Yale J Biol Med*. 2011;84:471-478.
3. Prasad V, Cifu A, Ioannidis JPA. Reversals of established medical practices: Evidence to abandon ship. *JAMA J Am Med Assoc*. 2012;307(1):37-38. doi:10.1001/jama.2011.1960.
4. Prasad V, Vandross A, Toomey C, et al. A decade of reversal: An analysis of 146 contradicted medical practices. *Mayo Clin Proc*. 2013;88(8):790-798. doi:10.1016/j.mayocp.2013.05.012.
5. Scott IA, Glasziou PP. Improving effectiveness of clinical medicine: The need for better translation of science into practice. *Med J Aust*. 2012;197(7):374-378. doi:10.5694/mja11.10365.
6. Marsh B. HRT "does more harm than good." *Daily Mail*. September 20, 2002.
7. Kingsley D. Some HRT does more harm than good. *ABC News*. July 10, 2002.
8. Ioannidis JPA. Contradicted and initially stronger effects in highly cited clinical research. *JAMA J Am Med Assoc*. 2005;294(2):218-228. doi:10.1001/jama.294.2.218.
9. Boden WE, O'Rourke RA, Teo KK, et al. Optimal medical therapy with or without PCI for Stable Coronary Disease. *N Engl J Med*. 2007;356(15):1503-1516.
10. Jensen ME, Evans AJ, Mathis JM, Kallmes DF, Cloft HJ, Dion JE. Percutaneous polymethylmethacrylate vertebroplasty in the treatment of osteoporotic vertebral body compression fractures: Technical aspects. *Am J Neuroradiol*. 1997;18:1897-1904.
11. Gray DT, Hollingworth W, Onwudiwe N, Beyo RA, Jarvik JG. Thoracic and lumbar vertebroplasties performed in US Medicare enrollees, 2001-2005. *J Am Med Assoc*. 2007;298(15):1760-1762. doi:10.1093/ageing/afp226.
12. Buchbinder R, Osborne R, Ebeling P, et al. A randomized trial of vertebroplasty for painful osteoporotic vertebral fractures. *N Engl J Med*. 2009;361(6):557-568. doi:10.1056/NEJMoa1201637.
13. Kallmes DF, Comstock B a, Heagerty PJ, et al. A randomized trial of vertebroplasty for osteoporotic spinal fractures. *N Engl J Med*. 2009;361(6):569-579. doi:10.1056/NEJMoa0900563.
14. Krieger N, Löwy I, Aronowitz R, et al. Hormone replacement therapy, cancer, controversies, and women's health: historical, epidemiological, biological, clinical, and advocacy perspectives. *J Epidemiol Community Health*. 2005;59:740-748. doi:10.1136/jech.2005.033316.
15. Women's Health Initiative Investigators Writing Group. Risks and Benefits of Estrogen Plus Progestin in Healthy Postmenopausal Women. *JAMA*. 2002;288(3):321-333.
16. Wilson RA, Wilson TA. The fate of the nontreated postmenopausal woman: A plea for the maintenance of adequate estrogen from puberty to the grave. *J Am Geriatr Soc*. 1963;11(4):347-362.
17. Rhoades FP. Minimizing the menopause. *J Am Geriatr Soc*. 1967;15(4):346-354.

18. Gallagher JC. Effect of early menopause on bone mineral density and fractures. *Menopause*. 2007;14(3):567-571. doi:10.1097/gme.0b013e31804c793d.
19. Francucci CM, Romagni P, Camilletti A, et al. Effect of natural early menopause on bone mineral density. *Maturitas*. 2008;59:323-328. doi:10.1016/j.maturitas.2008.03.008.
20. Hulley S, Grady D, Bush T, et al. Randomized Trial of Estrogen Plus Progestin for Secondary Prevention of Coronary Heart Disease in Postmenopausal Women. *JAMA J Am Med Assoc*. 1998;280(7):605. doi:10.1001/jama.280.7.605.
21. Gupta YK, Meenu M, Mohan P. The Tamiflu fiasco and lessons learnt. *Indian J Pharmacol*. 2015;47(1):11-16. doi:10.4103/0253-7613.150308.
22. Prasad VK, Cifu AS. *Ending Medical Reversal: Improving Outcomes, Saving Lives*. 1st ed. Baltimore, Maryland, USA: Johns Hopkins University Press; 2015.
23. Elliott RL. "Evidence-debased medicine" and the integrity of the medical profession. *J Clin Ethics*. 2011;22(1):71-73.
24. Boulware LE, Cooper LA, Ratner LE, LaVeist TA, Powe NR. Race and trust in the health care system. *Public Health Rep*. 2003;118(July-August):358-365. doi:10.1093/phr/118.4.358.
25. Sackett DL, Rosenberg WM, Gray J, Haynes RB, Richardson WS. Evidence based medicine: What it is and what it isn't. *BMJ*. 1996;312:71-72. doi:10.1136/bmj.312.7023.71.
26. Prasad V, Cifu A. A medical burden of proof: Towards a new ethic. *Biosocieties*. 2012;7(1):72-87. doi:10.1057/biosoc.2011.25.
27. Elshaug AG, Hiller JE, Tunis SR, Moss JR. Challenges in Australian policy processes for disinvestment from existing, ineffective health care practices. *Aust New Zealand Health Policy*. 2007;4:23. doi:10.1186/1743-8462-4-23.
28. Haas M, Hall J, Viney R, Gallego G. Breaking up is hard to do: Why disinvestment in medical technology is harder than investment. *Aust Heal Rev*. 2012;36:148-152. doi:10.1071/AH11032.
29. Robert G, Harlock J, Williams I. Disentangling rhetoric and reality: an international Delphi study of factors and processes that facilitate the successful implementation of decisions to decommission healthcare services. *Implement Sci*. 2014;9(123):1-15. doi:10.1186/s13012-014-0123-y.
30. Makic MBF, Rauen C, Watson R, Poteet AW. Examining the evidence to guide practice: Challenging practice habits. *Crit Care Nurse*. 2014;34(2):28-45. doi:10.4037/ccn2014262.
31. Rauen CA, Chulay M, Bridges E, Vollman KM, Arbour R. Seven evidence-based practice habits: putting some sacred cows out to pasture. *Crit Care Nurse*. 2008;28(2):98. doi:10.1017/CBO9781107415324.004.
32. Makic MBF, VonRueden KT, Rauen CA, Chadwick J. Evidence-based practice habits: Putting more sacred cows out to pasture. *Crit Care Nurse*. 2011;31(2):38-62. doi:10.4037/ccn2011908.
33. Makic MBF, Martin SA, Burns S, Philbrick D, Rauen C. Putting evidence into nursing practice : Four traditional practices not supported by the evidence. *Crit Care Nurse*. 2013;33(2):28-43. doi:10.4037/ccn2013787.
34. Wootton SH, Evans PW, Tyson JE. Unproven therapies in clinical research and practice: the necessity to change the regulatory paradigm. *Pediatrics*.

References

2013;132(4):599-601. doi:10.1542/peds.2013-0778.
35. Wellbery C, McAteer R. When medicine reverses itself: avoiding practice pitfalls. *Am Fam Physician*. 2013;88(11):737-738.
36. Fatovich DM. Medical reversal: What are you doing wrong for your patient today? *EMA - Emerg Med Australas*. 2013;25:1-3. doi:10.1111/1742-6723.12044.
37. Martin J, Cheng D. The real cost of care: focus on value for money, rather than price-tags. *Can J Anesth*. 2015;62:1034-1041. doi:10.1007/s12630-015-0444-6.
38. Chan AW, Song F, Vickers A, et al. Increasing value and reducing waste: Addressing inaccessible research. *Lancet*. 2014;383:257-266. doi:10.1016/S0140-6736(13)62296-5.
39. Ioannidis JPA. Why most published research findings are false. *PLoS Med*. 2005;2(8):0696-0701. doi:10.1371/journal.pmed.0020124.
40. Sumner P, Vivian-Griffiths S, Boivin J, et al. The association between exaggeration in health related science news and academic press releases: retrospective observational study. *Bmj*. 2014;349:g7015. doi:10.1136/bmj.g7015.
41. Goldacre B. Preventing bad reporting on health research. *BMJ*. 2014;349:g7465. doi:10.1136/bmj.f3817.
42. Ioannidis JPA. How many contemporary medical practices are worse than doing nothing or doing less? *Mayo Clin Proc*. 2013;88(8):779-781. doi:10.1016/j.mayocp.2013.05.010.
43. Montini T, Graham ID. "Entrenched practices and other biases": unpacking the historical, economic, professional, and social resistance to de-implementation. *Implement Sci*. 2015;10(24):1-8. doi:10.1186/s13012-015-0211-7.
44. Prasad V, Ioannidis JP. Evidence-based de-implementation for contradicted, unproven, and aspiring healthcare practices. *Implement Sci*. 2014;9(1):1-5. doi:10.1186/1748-5908-9-1.
45. Sutton D. Evidence Reversal: When New Evidence Contradicts Established Practices. 2015. http://ir.lib.uwo.ca/etd/3468.
46. Tatsioni A, Bonitsis NG, Ioannidis JP a. Persistence of contradicted claims in the literature. *JAMA*. 2007;298(21):2517-2526. doi:10.1016/j.jemermed.2008.02.043.
47. Flaherty DK. The vaccine-autism connection: A public health crisis caused by unethical medical practices and fraudulent science. *Ann Pharmacother*. 2011;45(10):1302-1304. doi:10.1345/aph.1Q318.
48. Shea BJ, Grimshaw JM, Wells G a, et al. Development of AMSTAR: a measurement tool to assess the methodological quality of systematic reviews. *BMC Med Res Methodol*. 2007;7:10. doi:10.1186/1471-2288-7-10.
49. Shea BJ, Hamel C, Wells GA, et al. AMSTAR is a reliable and valid measurement tool to assess the methodological quality of systematic reviews. *J Clin Epidemiol*. 2009;62:1013-1020. doi:10.1016/j.jclinepi.2008.10.009.
50. Ioannidis JPA, Lau J. Evolution of treatment effects over time: Empirical insight from recursive cumulative metaanalyses. *Proc Natl Acad Sci United States Am*. 2001;98(3):831-836. doi:10.1126/science.132.3438.1488.
51. Ioannidis JPA, Trikalinos TA, Ntzani EE, Contopoulos-Ioannidis DG. Genetic associations in large versus small studies: An empirical assessment. *Lancet*. 2003;361:567-571. doi:10.1016/S0140-6736(03)12516-0.
52. Doust J, Del Mar C. Why do doctors use treatments that do not work? *BMJ*.

2004;328(February):474-475. doi:10.1136/bmj.328.7438.474.
53. Trikalinos TA, Churchill R, Ferri M, et al. Effect sizes in cumulative meta-analyses of mental health randomized trials evolved over time. *J Clin Epidemiol*. 2004;57:1124-1130. doi:10.1016/j.jclinepi.2004.02.018.
54. Ioannidis JPA, Trikalinos TA. Early extreme contradictory estimates may appear in published research: The Proteus phenomenon in molecular genetics research and randomized trials. *J Clin Epidemiol*. 2005;58:543-549. doi:10.1016/j.jclinepi.2004.10.019.
55. Ioannidis JPA. Molecular bias. *Eur J Epidemiol*. 2005;20(9):739-745. doi:10.1007/s10654-005-2028-1.
56. Ioannidis JP. Evolution and translation of research findings: from bench to where? *PLoS Clin Trials*. 2006;1(7):e36. doi:10.1371/journal.pctr.0010036.
57. Shojania KG, Sampson M, Ansari MT, Ji J, Doucette S, Moher D. How quickly do systematic reviews go out of date ? A survival analysis. *Ann Intern Med*. 2007;147:224-233. doi:10.7326/0003-4819-147-4-200708210-00179.
58. Elshaug AG. Building the evidence base for disinvestment from ineffective health care practices: A case study in obstructive sleep apnoea syndrome. 2007;(October).
59. Ioannidis JPA. Non-replication and inconsistency in the genome-wide association setting. *Hum Hered*. 2007;64:203-213. doi:10.1159/000103512.
60. Elshaug AG, Moss JR, Littlejohns P, Karnon J, Merlin TL, Hiller JE. Identifying existing health care services that do not provide value for money. *Med J Aust*. 2009;190(5):269-273.
61. Tricoci P, Allen JM, Kramer JM, Califf RM, Smith Jr SC. Scientific evidence underlying the ACC / AHA clinical practice guidelines. *J Am Med Assoc*. 2009;301(8):831-841. doi:10.1001/jama.2009.205.
62. Thorlund K, Devereaux PJ, Wetterslev J, et al. Can trial sequential monitoring boundaries reduce spurious inferences from meta-analyses? *Int J Epidemiol*. 2009;38:276-286. doi:10.1093/ije/dyn179.
63. McCandless D. Snake Oil Supplements ? *Inf is Beautiful*. 2010.
64. Ioannidis JPA, Panagiotou OA. Comparison of effect sizes associated with biomarkers reported in highly cited individual articles and in subsequent meta-analyses. *J Am Med Assoc*. 2011;305(201):2200-2210. doi:10.1001/jama.2011.713.
65. Australia CMF for MBS. Australian Government Department of Health and Ageing, Medicare “Comprehensive Management Framework” environmental scan (Australia). 2015. www.health.gov.au/internet/main/publishing.nsf/Content/ReviewsCMFM. Accessed July 22, 2014.
66. Elshaug AG, Watt AM, Mundy L, Willis CD. Over 150 potentially low-value health care practices: an Australian study. *Med J Aust*. 2012;197(10):556-560. doi:10.5694/mja13.10080.
67. Ebell MH, Grad R. Top 20 research studies of 2011 for primary care physicians. *Am Fam Physician*. 2012;86(9):835-840.
68. Ray IB. Advancing evidence-based practice: A quarterly compilation of research updates most likely to change clinical practice. Winter 2012. *Ochsner J*. 2012;12(4):294-297.

69. Ray IB. Advancing evidence-based practice: A quarterly compilation of research updates most likely to change clinical practice. Fall 2012. *Ochsner J*. 2012;12(3):185-187.
70. Advancing Medical Professionalism to Improve Health Care Foundation. The American Board of Internal Medicine (ABIM) Foundation Initiative Choosing Wisely. www.choosingwisely.org/. Accessed July 22, 2014.
71. Polisena J, Clifford T, Mitton C, Elshaug AG, Russell E, Skidmore B. Case studies that illustrate disinvestment and resource allocation decision-making processes in health care: A systematic review. *Int J Technol Assess Health Care*. 2013;29(2):174-184. doi:10.1017/S0266462313000068.
72. Venkatesh AK, Schuur JD. A "top Five" list for emergency medicine: A policy and research agenda for stewardship to improve the value of emergency care. *Am J Emerg Med*. 2013;31:1520-1524. doi:10.1016/j.ajem.2013.07.019.
73. Kotzeva A, Torrente E, Almazán C, et al. Essencial : Adding value to healthcare through discontinuation of low-value practices. In: *2nd Conference of International Society for EBHC 6th International Conference for EBHC Teachers and Developers*. Taormina, Italy; 2013.
74. Scott IA, Elshaug AG. Foregoing low-value care: How much evidence is needed to change beliefs? *Intern Med J*. 2013;43:107-109. doi:10.1111/imj.12065.
75. Elshaug AG, Mcwilliams JM, Landon BE. The value of low-value lists. *JAMA J Am Med Assoc*. 2013;309(8):775-776. doi:10.1001/jama.2013.828.
76. Garner S, Docherty M, Somner J, et al. Reducing ineffective practice: challenges in identifying low-value health care using Cochrane systematic reviews. *J Health Serv Res Policy*. 2013;18(1):6-12. doi:10.1258/jhsrp.2012.012044.
77. Prasad V, Cifu A. The reversal of cardiology practices: interventions that were tried in vain. *Cardiovasc Diagn Ther*. 2013;3(4):228-235. doi:10.3978/j.issn.2223-3652.2013.10.05.
78. British Medical Journal. BMJ's Too Much Medicine. www.bmj.com/too-much-medicine. Accessed July 22, 2014.
79. Loder E, Weizenbaum E, Frishberg B, Silberstein S. Choosing wisely in headache medicine: The american headache society's list of five things physicians and patients should question. *Headache*. 2013;53:1651-1659. doi:10.1111/head.12233.
80. Ebell MH, Grad R. Top 20 research studies of 2012 for primary care physicians. *Am Fam Physician*. 2013;88(6):380-386.
81. Ray IB. Advancing evidence-based practice: A quarterly compilation of research updates most likely to change clinical practice. Spring 2013. *Ochsner J*. 2013;13(1):3-7.
82. Ray IB. Advancing evidence-based practice: A quarterly compilation of research updates most likely to change clinical practice. Fall 2013. *Ochsner J*. 2013;13(3):288-292.
83. Ray IB. Advancing evidence-based practice: A quarterly compilation of research updates most likely to change clinical practice. Winter 2013. *Ochsner J*. 2013;13(4):478-480.
84. Ray IB. Advancing evidence-based practice: A quarterly compilation of research updates most likely to change clinical practice. Summer 2013. *Ochsner J*. 2013;13(2):176-180.

References

85. McCandless D. Snake Oil Superfoods? *Inf is Beautiful*. 2013. www.informationisbeautiful.net/visualizations/snake-oil-superfoods/.
86. Hampton T. Clinical trial results may lead to changes in cardiovascular care. *JAMA - J Am Med Assoc*. 2014;312(19):1957-1959. doi:10.1001/jama.2014.14319.
87. Brien S, Gheihman G, Tse YK, Brynes M, Harrison S, Dobrow MJ. A scoping review of appropriateness of care research activity in Canada from a health system-level perspective. *Healthc Policy*. 2014;9(4):48-61. doi:10.12927/hcpol.2014.23773.
88. Bryson GL. Back to the future: Medical reversals and perioperative medicine. *Can J Anesth*. 2014;61:215-219. doi:10.1007/s12630-013-0103-8.
89. Ebell MH, Grad R. Top 20 research studies of 2013 for primary care physicians. *Am Fam Physician*. 2014;90(6):397-402.
90. McCandeless D. Snake Oil version 2. *Inf is Beautiful*. 2014. www.informationisbeautiful.net/2011/snake-oil-version-2/.
91. Ray IB. Advancing evidence-based practice: A quarterly compilation of research updates most likely to change clinical practice. Spring 2014. *Ochsner J*. 2014;14(1):3-6.
92. Ray IB. Advancing evidence-based practice: A quarterly compilation of research updates most likely to change clinical practice. Summer 2014. *Ochsner J*. 2014;14(2):148-153.
93. Ray IB. Advancing evidence-based practice: A quarterly compilation of research updates most likely to change clinical practice. Winter 2014. *Ochsner J*. 2014;14(4):521-526.
94. Macleod MR, Michie S, Roberts I, et al. Biomedical research: Increasing value, reducing waste. *Lancet*. 2014;383:101-104. doi:10.1016/S0140-6736(13)62329-6.
95. Choosing Wisely Canada. The Canadian Medical Association (CMA) Campaign Choosing Wisely. www.choosingwiselycanada.org/. Accessed July 22, 2014.
96. Gnjidic D, Elshaug AG. De-adoption and its 43 related terms: harmonizing low-value care terminology. *BMC Med*. 2015;13(1):273. doi:10.1186/s12916-015-0511-4.
97. Niven DJ, Mrklas KJ, Holodinsky JK, et al. Towards understanding the de-adoption of low-value clinical practices: a scoping review. *BMC Med*. 2015;13:255. doi:10.1186/s12916-015-0488-z.
98. Paprica PA, Culyer AJ, Elshaug AG, Peffer J, Sandoval GA. From talk to action: Policy stakeholders, appropriateness, and selective disinvestment. *Int J Technol Assess Health Care*. 2015;31(4):236-240. doi:10.1017/S0266462315000392.
99. Mayer J, Nachtnebel A. Disinvesting from ineffective technologies: Lessons learned from current programs. *Int J Technol Assess Health Care*. 2015;31(6):355-362. doi:10.1017/s0266462315000641.
100. Mitera G, Earle C, Latosinsky S, et al. Choosing Wisely Canada cancer list : Ten low-value or harmful practices that should be avoided in cancer care. *J Oncol Pract*. 2015;11(3):e296-e303. doi:10.1200/JOP.2015.004325.
101. Duckett SJ, Breadon P, Romanes D. Identifying and acting on potentially inappropriate care. *Med J Aust*. 2015;203(4):1-6. doi:10.5694/mja15.01241.
102. Selby K, Gaspoz J-M, Rhodondi N, et al. Creating a list of low-value health care activities in swiss primary care. *JAMA J Am Med Assoc*. 2015;175(4):640-642.

References

doi:10.1001/jamainternmed.2014.8020.
103. Cifu AS, Prasad VK. Medical debates and medical reversal. *J Gen Intern Med*. 2015;30(12):1729-1730. doi:10.1007/s11606-015-3481-5.
104. Laiteerapong N, Huang ES. The pace of change in medical practice and health policy: Collision or coexistence? *J Gen Intern Med*. 2015;30(6):848-852. doi:10.1007/s11606-015-3182-0.
105. Wang MTM, Gamble G, Grey A. Letter: responses of specialist societies to evidence for reversal of practice. *JAMA Intern Med*. 2015;175(5):845-848. doi:10.1001/jamainternmed.2015.0153.
106. Malhotra A, Maughan D, Ansell J, et al. Choosing Wisely in the UK: The Academy of Medical Royal Colleges' initiative to reduce the harms of too much medicine. *BMJ*. 2015;350:h2308. doi:10.1136/bmj.h2308.
107. Morgan D, Wright S, Dhruva S. Update on medical overuse. *JAMA J Am Med Assoc*. 2015;175(1):120-124. doi:10.1001/jamainternmed.2014.5444.
108. Ebell MH, Grad R. Top 20 research studies of 2014 for primary care physicians. *Am Fam Physician*. 2015;92(5):377-383.
109. Sundsted KK, Wieland ML, Szostek JH, Post JA, Mauck KF. Update in outpatient general internal medicine : Practice-changing evidence published in 2014. *Am J Med*. 2015;128(10):1065-1069. doi:10.1016/j.amjmed.2015.04.033.
110. Finn KM, Greenwald JL. Update in hospital medicine: Evidence you should know. *J Hosp Med*. 2015;10(12):817-826. doi:10.1002/jhm.2476.
111. Hanrahan K, Wagner M, Matthews G, et al. Sacred cow gone to pasture: A systematic evaluation and integration of evidence-based practice. *Worldviews Evidence-Based Nurs*. 2015;12(1):3-11. doi:10.1111/wvn.12072.
112. Davidoff F. On the undiffusion of established practices. *JAMA Intern Med*. 2015;175(5):809-811. doi:10.1001/jamainternmed.2015.0167.
113. Nottinghamshire Healthcare, NICE. *NICE "Do Not Do" Recommendations*. Vol 9.; 2007.
114. US Preventative Services Task Force (USPSTF). U.S. Preventive Services Task Force (USPSTF) "Grade 'D' recommendations" for preventive health services. 2016;(May). www.uspreventiveservicestaskforce.org/. Accessed July 22, 2016.
115. Sprenger M, Robausch M, Moser A. Quantifying low-value services by using routine data from Austrian primary care. *Eur J Public Health*. 2016:80. doi:10.1093/eurpub/ckw080.
116. National Institute for Health and Care (NICE). UK Database of Uncertainties about the Effects of Treatments (UK DUETs). http://www.library.nhs.uk/duets/. Accessed July 22, 2014.
117. Drazer MW, Salama JK, Hahn OM, Weichselbaum RR, Chmura SJ. Stereotactic body radiotherapy for oligometastatic breast cancer: a new standard of care, or a medical reversal in waiting? *Expert Rev Anticancer Ther*. 2016;16(6):625-632. doi:10.1080/14737140.2016.1178577.
118. Prasad V. Translation failure and medical reversal: Two sides to the same coin. *Eur J Cancer*. 2016;52:197-200. doi:10.1016/j.ejca.2015.08.024.
119. Ebell MH, Grad GR. Top 20 research studies of 2015 for primary care physicians. *Am Fam Physician*. 2016;93(1):756-762.
120. Szostek JH, Wieland ML, Post JA, Sundsted KK, Mauck KF. Update in outpatient

general internal medicine: Practice-changing evidence published in 2015. *Am J Med*. 2016. doi:10.1016/j.amjmed.2016.03.004.
121. Singh N, Gupta M. Impactful clinical trials of 2015: What clinicians need to know. *Can J Cardiol*. 2016;0(0). doi:10.1016/j.cjca.2013.03.003.
122. Sutton D, Qureshi R, Martin J. Evidence reversal – when new evidence contradicts current claims : A systematic overview review. *Submitted*. 2016.
123. Balshem H, Helfand M, Schunemann HJ, et al. GRADE guidelines: 3. Rating the quality of evidence. *J Clin Epidemiol*. 2011;64:401-406. doi:10.1016/j.jclinepi.2010.07.015.
124. GOOGLE. Top Publications in Health & Medical Sciences. *5-year Hirsch Index*. 2015. https://scholar.google.ca/citations?view_op=top_venues&hl=en&vq=med. Accessed March 8, 2017.
125. Higgins J, Green S, eds. *Cochrane Handbook for Systematic Reviews of Interventions*. 5.1.0.; 2011. www.handbook.cochrane.org.
126. Guyatt GH, Oxman AD, Schünemann HJ, Tugwell P, Knottnerus A. GRADE guidelines: a new series of articles in the Journal of Clinical Epidemiology. *J Clin Epidemiol*. 2011;64:380-382. doi:10.1016/j.jclinepi.2010.09.011.
127. Guyatt GH, Oxman AD, Sultan S, et al. GRADE guidelines: 9. Rating up the quality of evidence. *J Clin Epidemiol*. 2011;64(12):1311-1316. doi:10.1016/j.jclinepi.2011.06.004.
128. Vach W. *Regression Models as a Tool in Medical Research*. Boca Raton, Florida: Taylor and Francis Group; 2013.
129. StataCorp. Stepwise - Stepwise estimation. :1-10. http://www.stata.com/manuals13/rstepwise.pdf.
130. Harrell FE. *Regression Modeling Strategies: With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analyis*. Vol 64. 2nd ed. Springer; 2015. doi:10.1007/978-1-4757-3462-1.
131. Gail M, Krickeberg K, Samet JM, Tsiatis A, Wong W. *Regression Methods in Biostatistics*. 2nd ed. New York: Springer US; 2012.
132. Akl EA, Briel M, You JJ, et al. Potential impact on estimated treatment effects of information lost to follow-up in randomised controlled trials (LOST-IT): systematic review. *Bmj*. 2012;344:e2809. doi:10.1136/bmj.e2809.
133. Davis JS, He V, Anstey NM, Condon JR. Long term outcomes following hospital admission for sepsis using relative survival analysis: A prospective cohort study of 1,092 patients with 5 year follow up. *PLoS One*. 2014;9(12):e112224. doi:10.1371/journal.pone.0112224.
134. Vincent J-L. Endpoints in sepsis trials: more than just 28-day mortality? *Crit Care Med*. 2004;32(5 Suppl):S209-S213. doi:10.1097/01.CCM.0000126124.41743.86.
135. CONSORT Group. CONSORT abstract checklist. 2010.
136. Dwan K, Altman DG, Arnaiz JA, et al. Systematic review of the empirical evidence of study publication bias and outcome reporting bias. *PLoS One*. 2008;3(8):e3081. doi:10.1371/journal.pone.0003081.
137. Bhandari M, Busse JW, Jackowski D, et al. Association between industry funding and statistically significant pro-industry findings in medical and surgical randomized trials. *CMAJ*. 2004;170(4):477-480. doi:http://www.cmaj.ca/cgi/content/full/170/4/481.

138. Lexchin J. Those who have the gold make the evidence: How the pharmaceutical industry biases the outcomes of clinical trials of medications. *Sci Eng Ethics*. 2012;18:247-261. doi:10.1007/s11948-011-9265-3.
139. Naci H, Ioannidis JPA. How Good Is " Evidence " from Clinical Studies of Drug Effects and Why Might Such Evidence Fail in the Prediction of the Clinical Utility of Drugs? *Annu Rev Pharmacol Toxicol*. 2015;55:169-189. doi:10.1146/annurev-pharmtox-010814-124614.
140. Walsh M, Srinathan SK, McAuley DF, et al. The statistical significance of randomized controlled trial results is frequently fragile: A case for a Fragility Index. *J Clin Epidemiol*. 2014;67:622-628. doi:10.1016/j.jclinepi.2013.10.019.
141. De Smedt A, Buyl R, Nyssen M. Evidence-based practice in primary health care. *Stud Health Technol Inform*. 2006;124:651-656. http://search.ebscohost.com/login.aspx?direct=true&db=cmedm&AN=17108590&site=ehost-live.
142. Tenopir C, King D w, Bush A. Medical faculty's use of print and electronic journals: changes over time and in comparison with scientists. *J Med Libr Assoc*. 2004;92(2):233-241. http://search.ebscohost.com/login.aspx?direct=true&db=llf&AN=502926930&site=ehost-live%5Cnhttp://www.ncbi.nlm.nih.gov/pmc/articles/PMC385305/pdf/i0025-7338-092-02-0233.pdf.
143. McKibbon KA, Wilczynski NL, Haynes RB. What do evidence-based secondary journals tell us about the publication of clinically important articles in primary healthcare journals? *BMC Med*. 2004;2:33. doi:10.1186/1741-7015-2-33.
144. Lenzer J. Why we can't trust clinical guidelines. *BMJ*. 2013;346:f3830. doi:10.1136/bmj.f3830.
145. Kung J, Miller RR, Mackowiak PA. Failure of clinical practice guidelines to meet institute of medicine standards: two more decades of little, if any, progress. *Arch Intern Med*. 2012;172(21):1628-1633. doi:10.1001/2013.jamainternmed.56.
146. Chalmers I, Altman D, McHaffie H, Owens N, Cooke R. Data sharing among data monitoring committees and responsibilities to patients and science. *Trials*. 2013;14(1):102. doi:10.1186/1468-6708-14-102.
147. Villas Boas PJF, Spagnuolo RS, Kamegasawa A, et al. Systematic reviews showed insufficient evidence for clinical practice in 2004: What about in 2011? the next appeal for the evidence-based medicine age. *J Eval Clin Pract*. 2013;19:633-637. doi:10.1111/j.1365-2753.2012.01877.x.
148. Runciman WB, Hunt TD, Hannaford NA, et al. CareTrack: Assessing the appropriateness of health care delivery in Australia. *Med J Aust*. 2012;197(2):100-105. doi:10.5694/mja12.10510.
149. Kostkova P, Brewer H, de Lusignan S, et al. Who owns the data? Open data for healthcare. *Front Public Heal*. 2016;4:1-7. doi:10.3389/fpubh.2016.00007.
150. Hunter A, Williams M. Aggregating evidence about the positive and negative effects of treatments. *Artif Intell Med*. 2012;56:173-190. doi:10.1016/j.artmed.2012.09.004.
151. Guallar E, Laine C. Controversy over clinical guidelines: Listen to the evidence, not the noise. *Ann Intern Med*. 2014;160(5):361-362. doi:10.7326/M14-0112.

152. Doherty S. History of evidence-based medicine. Oranges, chloride of lime and leeches: barriers to teaching old dogs new tricks. *Emerg Med Australas*. 2005;17:314-321. doi:10.1111/j.1742-6723.2005.00752.x.
153. Grol R, Grimshaw J. From best evidence to best practice: Effective implementation of change in patients' care. *Lancet*. 2003;362:1225-1230. doi:10.1016/S0140-6736(03)14546-1.
154. Cabana MD, Rand CS, Powe NR, et al. Why don't physicians follow clinical practice guidelines? A framework for improvement. *JAMA J Am Med Assoc*. 1999;Vol 282(15):1458-1465. doi:10.1001/jama.282.15.1458.
155. Grol R. Successes and Failures in the Implementation of Evidence-Based Guidelines for Clinical Practice. *Med Care*. 2001;39(8):II-46-II-54.
156. McGlynn E a., Asch SM, Adams J, et al. The quality of health care delivered to adults in the United States. *N Engl J Med*. 2003;348(26):2635-2645. http://www.ncbi.nlm.nih.gov/pubmed/14606462.
157. Ioannidis JPA. Why most discovered true associations are inflated. *Epidemiology*. 2008;19(5):640-648. doi:10.1097/EDE.ObO.
158. National Institute for Health and Care (NICE). National Institute for Health and Care Excellence (NICE) "Do not do" list. 2007. www.nice.org.uk/savingsandproductivity/collection?page=1&pagesize=2000&type=do not do. Accessed July 22, 2014.
159. Muellerleile P, Mullen B. Sufficiency and stability of evidence for public health interventions using cumulative meta-analysis. *Am J Public Health*. 2006;96(3):515-522. doi:10.2105/AJPH.2003.036343.
160. Dent L, Taylor R, Jolly K, Raftery J. "Flogging dead horses": evaluating when have clinical trials achieved sufficiency and stability? A case study in cardiac rehabilitation. *Trials*. 2011;12:83. doi:10.1186/1745-6215-12-83.

APPENDICES

APPENDIX A
SYSTEMATIC REVIEW METHODOLOGY

DATABASE SEARCH STRATEGIES & PRISMA FLOW DIAGRAM

Table 1: Database Search Strategies and Results

PUBMED Database Search Strategy and Results (July 6th, 2016)

Search	Search Terms	Articles
1	("Evidence-based practice"[MeSH Major Topic] OR "Patient care management"[MeSH Major Topic]) OR "guidelines as topic"[MeSH Major Topic]	417,307
2	((((“clinical practice”[All Fields] OR “practice guideline*”[All Fields]) OR “physician's practice pattern”[All Fields]) OR “evidence-based”[All Fields]) OR “evidence based”[All Fields])	259,320
3	1 or 2	626,571
4*	((((((((((publication) OR publish*) OR evidence) OR practice) OR guideline*) OR medical) OR clinical) OR standard) OR standards) OR unexpected) OR surprising) N3 ((((((revers*) OR change) OR contradict*) OR divest*) OR de-implement*))[All Fields] OR disinvest*[All Fields]	351
5	3 and 4	55
6	prasad v[Author] OR ioannidis j[Author] OR ioannidis jp[Author] OR cifu a[Author] OR elshaug a[Author]	1,804
7	5 or 6	1,844
FINAL	Limit 7 to yr=“2014-Current”	443

MEDLINE (EMBASE) Database Search Strategy and Results (July 1st, 2016)

Search	Search Terms	Articles
1	exp Evidence-Based Practice/ OR exp Patient Care Management/ OR exp Guidelines as Topic/	777,823
2	(clinical practice OR practice guideline* OR physician's practice pattern OR evidence-based OR evidence based)	331,623
3	1 or 2	937,017
4*	((publication OR publish* OR evidence OR practice OR guideline* OR medical OR clinical OR standard OR standards OR unexpected OR surprising) adj3 (revers* OR change OR contradict* OR divest* OR de-implement*)).mp. OR disinvest*.mp.	19,891
5	3 and 4	4,577
6	(prasad v OR ioannidis j OR ioannidis jp OR cifu a OR elshaug a).au.	1,421
7	5 or 6	4,577
FINAL	Limit 7 to yr=“2014-Current”	957

MEDLINE (OVID) Database Search Strategy and Results (July 22nd, 2014)

Search	Search Terms	Articles
1	exp Evidence-Based Practice/ OR exp Patient Care Management/ OR exp Guidelines as Topic/	669,457
2	(clinical practice OR practice guideline* OR physician's practice pattern OR evidence-based OR evidence based)	273,895
3	1 or 2	794,592
4*	((publication OR publish* OR evidence OR practice OR guideline* OR medical OR clinical OR standard OR standards OR unexpected OR surprising) adj3 (revers* OR change OR contradict* OR divest* OR de-implement*)).mp. OR disinvest*.mp.	16,538
5	3 and 4	3,676

6	(prasad v OR ioannidis j OR ioannidis jp OR cifu a OR elshaug a).au.	1,152
FINAL	5 or 6	4,811

* For search #4, the original search strategy was missing a space between "disinvest*" and "OR," thus the final search string lacked results generated from "OR disinvest* OR result*". The updated search strategy includes "disinest*" but excludes "result*" due to an extraneous number of results.

TOTAL ARTICLES RETRIEVED FROM MEDLINE	6,201

EMBASE (EMBASE) Database Search Strategy and Results (July 1st, 2016)

Search	Search Terms	Articles
1	exp Evidence-Based Practice/ OR exp Patient Care Management/ OR exp Guidelines as Topic/	1,704,315
2	(clinical practice OR practice guideline* OR physician's practice pattern OR evidence-based OR evidence based)	680,007
3	1 or 2	1,935,989
4*	((publication OR publish* OR evidence OR practice OR guideline* OR medical OR clinical OR standard OR standards OR unexpected OR surprising) adj3 (revers* OR change OR contradict* OR divest* OR de-implement*)).mp. OR disinvest*.mp.	27,833
5	3 and 4	7,960
6	(prasad v OR ioannidis j OR ioannidis jp OR cifu a OR elshaug a).au.	859
7	5 or 6	7,960
FINAL	Limit 7 to yr="2014-Current"	1,902

EMBASE (OVID) Database Search Strategy and Results (July 22nd, 2014)

Search	Search Terms	Articles
1	exp Evidence-Based Practice/ OR exp Patient Care Management/ OR exp Guidelines as Topic/	1,407,119
2	(clinical practice OR practice guideline* OR physician's practice pattern OR evidence-based OR evidence based)	566,343
3	1 or 2	1,598,004
4*	((publication OR publish* OR evidence OR practice OR guideline* OR medical OR clinical OR standard OR standards OR unexpected OR surprising) adj3 (revers* OR change OR contradict* OR divest* OR de-implement*)).mp. OR disinvest*.mp.	24,119
5	3 and 4	6,262
6	(prasad v OR ioannidis j OR ioannidis jp OR cifu a OR elshaug a).au.	637
FINAL	5 or 6	6,894

* For search #4, the original search strategy was missing a space between "disinvest*" and "OR," thus the final search string lacked results generated from "OR disinvest* OR result*". The updated search strategy includes "disinest*" but excludes "result*" due to an extraneous number of results.

TOTAL ARTICLES RETRIEVED FROM EMBASE	8,796

CINAHL Database Search Strategy and Results (July 1st, 2016)

Search	Search Terms	Articles
1	(MH "Professional Practice, Evidence-Based") OR (MH "Evidence- Based Dental Practice") OR (MH "Medical Practice, Evidence- Based") OR (MH "Nursing Practice, Evidence-Based") OR (MH "Occupational Therapy Practice, Evidence-Based") OR (MH "Physical Therapy Practice, Evidence-Based") OR (MH "Professional Practice, Research- Based") OR (MH "Physical Therapy Practice, Research-Based") OR (MH "Occupational Therapy Practice, Research-Based") OR (MH "Practice Guidelines") OR (MH "Nursing Practice, Research-Based") OR (MH "Medical Practice, Research- Based") OR (MH "Practice Patterns") OR (MH "Medical Practice") OR (MH "Nursing Care Plans")	76,539
2	(clinical practice OR practice guideline* OR physician's practice pattern OR evidence- based OR evidence based)	126,752
3	1 or 2	139,840
4	((publication OR publish* OR evidence OR practice OR guideline* OR medical OR clinical OR standard OR standards OR surprising OR unexpected) N3 (revers* OR change OR contradict* OR divest* OR de-implement* OR disinvest* OR result*))	40,459
5	3 and 4	7,653
6	AU (prasad v OR ioannidis j OR ioannidis jp OR ioannidis jpa OR cifu a OR elshaug a)	190
7	5 or 6	7,839
FINAL	Limiters – Published date: 20140101-20161231	1,250

CINAHL Database Search Strategy and Results (July 22nd, 2014)

Search	Search Terms	Articles
1	(MH "Professional Practice, Evidence-Based") OR (MH "Evidence-Based Dental Practice") OR (MH "Medical Practice, Evidence-Based") OR (MH "Nursing Practice, Evidence-Based") OR (MH "Occupational Therapy Practice, Evidence-Based") OR (MH "Physical Therapy Practice, Evidence-Based") OR (MH "Professional Practice, Research- Based") OR (MH "Physical Therapy Practice, Research-Based") OR (MH "Occupational Therapy Practice, Research-Based") OR (MH "Practice Guidelines") OR (MH "Nursing Practice, Research-Based") OR (MH "Medical Practice, Research-Based") OR (MH "Practice Patterns") OR (MH "Medical Practice") OR (MH "Nursing Care Plans")	106,511
2	(clinical practice OR practice guideline* OR physician's practice pattern OR evidence- based OR evidence based)	137,475
3	1 or 2	156,054
4	((publication OR publish* OR evidence OR practice OR guideline* OR medical OR clinical OR standard OR standards OR surprising OR unexpected) N3 (revers* OR change OR contradict* OR divest* OR de-implement* OR disinvest* OR result*))	36,042
5	3 and 4	6,313
6	AU (prasad v OR ioannidis j OR ioannidis jp OR ioannidis jpa OR cifu a OR elshaug a)	268
FINAL	5 or 6	6,577

TOTAL ARTICLES RETRIEVED FROM CINAHL	7,827

Web of Science Database Search Strategy and Results (July 3rd, 2016)

Search	Search Terms	Articles
1	TS=(Evidence-Based Practice OR Patient Care Management OR Guidelines)	447,145
2	TI=(clinical practice OR practice guideline* OR physician's practice pattern OR evidence- based OR evidence based)	61,470
3	1 or 2	487,183
4	TI=((publication OR publish* OR evidence OR practice OR guideline* OR medical OR clinical OR standard OR standards OR surprising OR unexpected) NEAR/3 (revers* OR change OR contradict* OR divest* OR de-implement* OR disinvest* OR result*))	30,851
5	3 and 4	2,937
6	AU=(Prasad V OR Ioannidis J OR Ioannidis JP OR Ioannidis JPA OR Cifu A OR Elshaug A)	3,502
7	5 or 6	6,435
FINAL	#7 From 2014-2016	1,215

Web of Science Database Search Strategy and Results (July 22nd, 2014)

Search	Search Terms	Articles
1	TS=(Evidence-Based Practice OR Patient Care Management OR Guidelines)	354,556
2	TI=(clinical practice OR practice guideline* OR physician's practice pattern OR evidence- based OR evidence based)	51,412
3	1 or 2	388,323
4	TI=((publication OR publish* OR evidence OR practice OR guideline* OR medical OR clinical OR standard OR standards OR surprising OR unexpected) NEAR/3 (revers* OR change OR contradict* OR divest* OR de-implement* OR disinvest* OR result*))	27,607
5	3 and 4	2,398
6	AU=(Prasad V OR Ioannidis J OR Ioannidis JP OR Ioannidis JPA OR Cifu A OR Elshaug A)	2,990
FINAL	5 or 6	5,384

TOTAL ARTICLES RETRIEVED FROM WEB OF SCIENCE	6,599

Dissertations and Theses Database Search Strategy and Results (July 3rd, 2016)

Search	Search Terms	Articles
1	su(Evidence-Based Practice OR Patient Care Management OR Guidelines)	1,816
2	all(clinical practice OR practice guideline* OR physician's practice pattern OR evidence- based OR evidence based)	88,179
3	1 or 2	89,262
4	all((publication OR publish* OR evidence OR practice OR guideline* OR medical OR clinical OR standard OR standards OR surprising OR unexpected) NEAR/3 (revers* OR change OR contradict* OR divest* OR de-implement* OR disinvest* OR result*))	42,744
5	3 and 4	6,607
FINAL	Limit to 2014-2016	893

Dissertations and Theses Database Search Strategy and Results (July 22nd, 2014)

Search	Search Terms	Articles
1	su(Evidence-Based Practice OR Patient Care Management OR Guidelines)	1,245
2	all(clinical practice OR practice guideline* OR physician's practice pattern OR evidence- based OR evidence based)	73,768
3	1 or 2	74,504
4	all((publication OR publish* OR evidence OR practice OR guideline* OR medical OR clinical OR standard OR standards OR surprising OR unexpected) NEAR/3 (revers* OR change OR contradict* OR divest* OR de-implement* OR disinvest* OR result*))	9,517
FINAL	3 and 4	1,349

TOTAL ARTICLES RETRIEVED FROM DISSERTATIONS AND THESES	2,242

Canadian Health Research Collection (CHRC) Database Search Strategy and Results (July 5th, 2016)

Search	Search Terms	Articles
1	((Evidence-Based Practice OR Patient Care Management OR Guidelines) OR (clinical practice OR practice guideline* OR physician's practice pattern OR evidence-based OR evidence based)) AND ((publication OR publish* OR evidence OR practice OR guideline* OR medical OR clinical OR standard OR standards) WITHIN-3 (revers* OR change OR contradict* OR divest* OR de-implement* OR disinvest* OR surprising result*))	13,268
FINAL	Limit to 2014-2016	1,900

Canadian Health Research Collection (CHRC) Database Search Strategy and Results (July 22nd, 2014)

Search	Search Terms	Articles
1	All:(Evidence-Based Practice OR Patient Care Management OR Guidelines)	10,902
2	All:(clinical practice OR practice guideline* OR physician's practice pattern OR evidence- based OR evidence based)	10,734
3	1 or 2	11,316
4	All:((publication OR publish* OR evidence OR practice OR guideline* OR medical OR clinical OR standard OR standards) WITHIN-3 (revers* OR change OR contradict* OR divest* OR de-implement* OR disinvest* OR surprising result*))	11,553
FINAL	3 and 4	11,225

TOTAL ARTICLES RETRIEVED FROM CHRC	13,125

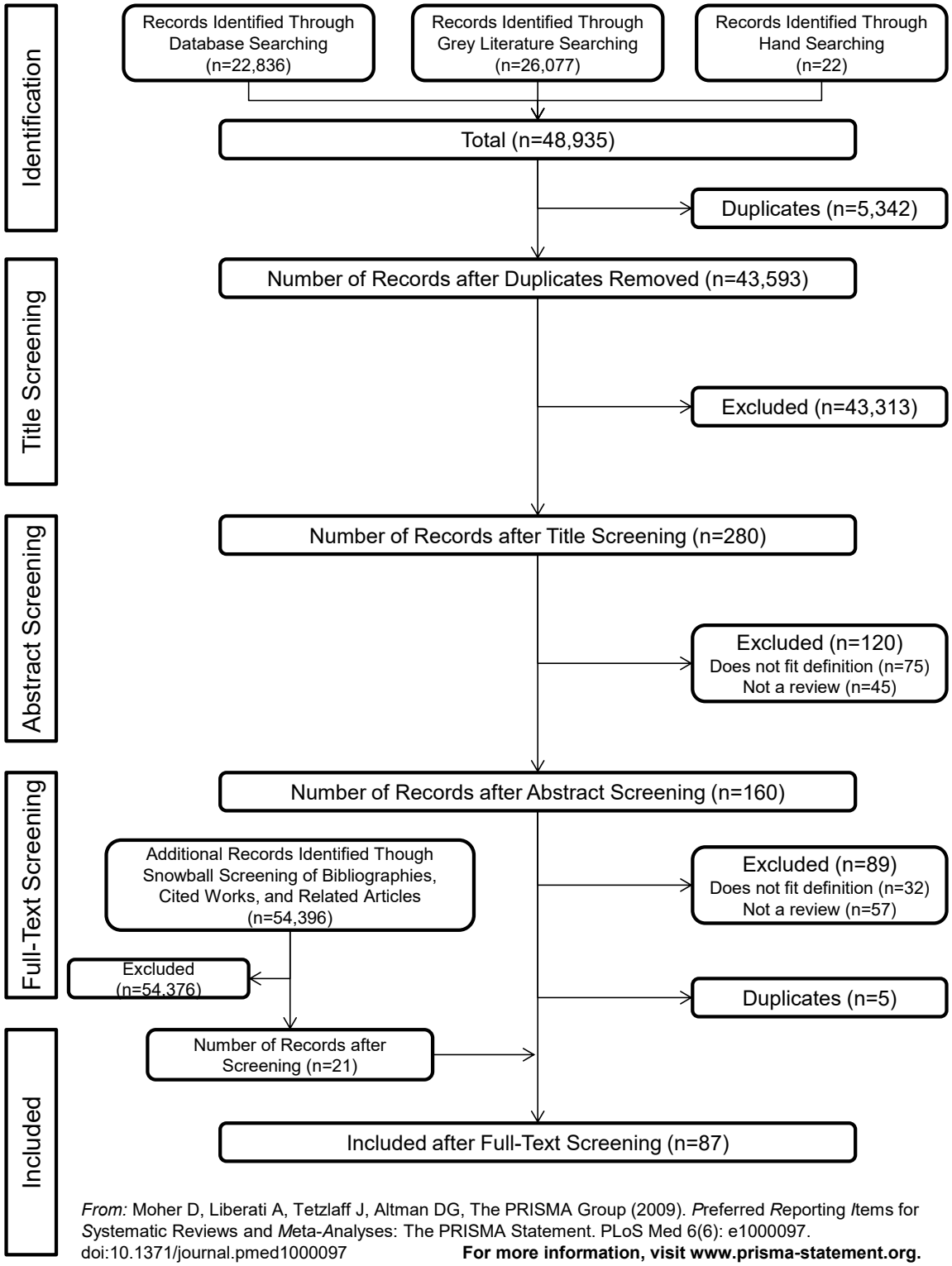
Google Scholar Database Search Strategy and Results (July 22nd, 2014)

Search	Search Terms	Articles
FINAL	Author Profile for "Vinay Prasad" [Custom Range: 2014-2016]	86
FINAL	Author Profile for "Adam Cifu" [Custom Range: 2014-2016]	32
FINAL	Author Profile for "John P. A. Ioannidis" [Custom Range: 2014-2016]	203
FINAL	Author Profile for "Dr Adam Elshaug" [Custom Range: 2014-2016]	16
FINAL	("evidence reversal" OR "clinical reversal" OR "medical reversal" OR "divestment" OR "de-implement" OR "disinvestment" OR "surprising result") AND ("evidence based" OR "evidence-based") [Custom Range: 2014-2016]	1,000
	Total	1,337

Google Scholar Database Search Strategy and Results (July 22nd, 2014)

Search	Search Terms	Articles
FINAL	Author Profile for "Vinay Prasad" [Custom Range: 2014-2016]	93
FINAL	Author Profile for "Adam Cifu" [Custom Range: 2014-2016]	32
FINAL	Author Profile for "John P. A. Ioannidis" [Custom Range: 2014-2016]	1,132
FINAL	Author Profile for "Dr Adam Elshaug" [Custom Range: 2014-2016]	90
FINAL	("evidence reversal" OR "clinical reversal" OR "medical reversal" OR "divestment" OR "de-implement" OR "disinvestment" OR "surprising result") AND ("evidence based" OR "evidence-based") [Custom Range: 2014-2016]	969
	Total	2,316

TOTAL ARTICLES RETRIEVED FROM GOOGLE SCHOLAR	3,653



aFigure 1: PRISMA Flow Diagram

APPENDIX B
DATA EXTRACTION FOR 87 INCLUDED ARTICLES

Table 2: Complete data extraction of 87 included articles

Author(s) or Organization	Title	Year	Study Type	Term(s) Used	Relationship to Reversal	AMSTAR Rating Score	AMSTAR Quality Rating
Advancing Medical Professionalism to Improve Health Care Foundation[1]	The American Board of Internal Medicine (ABIM) Foundation Initiative Choosing Wisely	2012	Systematic Review	Things providers and patients should question	Target	1	Very Low Quality
				Unnecessary medical tests, treatments and procedures	Target		
				Wasteful medical tests, treatments and procedures	Target		
Australia Comprehensive Management Framework for Medicare Benefits Schedule[2]	Australian Government Department of Health and Ageing, Medicare 'Comprehensive Management Framework' environmental scan (Australia)	2011	Collection of Studies	Recommendations for practice	Target	4	Low Quality
Brien et al[3]	A scoping review of appropriateness of care research activity in Canada from a health system-level perspective	2014	Systematic Scoping Review	Inappropriate care	Target	5	Low Quality
British Medical Journal[4]	BMJ's Too Much Medicine	2013	Systematic Review	Overdiagnosis	Phenomenon	1	Very Low Quality
				Too Much Medicine	Consequence		
Bryson[5]	Back to the future: Medical reversals and perioperative medicine	2014	Collection of Studies	Medical reversal	Phenomenon	1	Very Low Quality

Author(s) or Organization	Title	Year	Study Type	Term(s) Used	Relationship to Reversal	AMSTAR Rating Score	AMSTAR Quality Rating
Choosing Wisely Canada[6]	The Canadian Medical Association (CMA) Campaign Choosing Wisely	2014	Systematic Review	Unnecessary tests, treatments and procedures	Target	1	Very Low Quality
Cifu and Prasad[7]	Medical debates and medical reversal	2015	Collection of Studies	Medical reversal	Phenomenon	0	Very Low Quality
Davidoff[8]	On the undiffusion of established medical practices	2015	Collection of Studies	Undiffusion	Consequence	1	Very Low Quality
Doust and Del Mar[9]	Why do doctors use treatments that do not work?	2004	Collection of Studies	Ineffective or harmful interventions	Target	1	Very Low Quality
Drazer et al[10]	Stereostatic body radiotherapy for oligometastatic breast cancer: A new standard of care, or a medical reversal in waiting?	2016	Systematic Review	Medical reversal	Phenomenon	1	Very Low Quality
Duckett et al[11]	Identifying and acting on potentially inappropriate care	2015	Secondary Data Analysis and Review	Inappropriate care	Target	2	Very Low Quality
Ebell and Grad[12]	Top 20 research studies of 2011 for primary care physicians	2012	Collection of Studies	POEMs likely to change practice	Target	2	Very Low Quality
Ebell and Grad[13]	Top 20 research studies of 2012 for primary care physicians	2013	Collection of Studies	POEMs likely to change practice	Target	3	Low Quality
Ebell and Grad[14]	Top 20 research studies of 2013 for primary care physicians	2014	Collection of Studies	POEMs likely to change practice	Target	3	Low Quality
Ebell and Grad[15]	Top 20 research studies of 2014 for primary care physicians	2015	Collection of Studies	POEMs likely to change practice	Target	4	Low Quality

Author(s) or Organization	Title	Year	Study Type	Term(s) Used	Relationship to Reversal	AMSTAR Rating Score	AMSTAR Quality Rating
Ebell and Grad[16]	Top 20 research studies of 2015 for primary care physicians	2016	Collection of Studies	POEMs likely to change practice	Target	5	Low Quality
Elshaug et al[17]	Building the evidence base for disinvestment from ineffective health care practices: a case study in obstructive sleep apnoea syndrome	2007	Collection of Studies	Disinvestment	Consequence	2	Very Low Quality
Elshaug et al[18]	Identifying existing health care services that do not provide value for money	2009	Collection of Studies	Assess new interventions – displace old	Consequence	1	Very Low Quality
				Ineffective, harmful, or non-cost-effective interventions	Target		
				Legacy items	Target		
				Technology development	Target		
Elshaug et al[19]	Over 150 potentially low-value health care practices: an Australian study	2012	Systematic Review	Low value care	Target	5	Low Quality
Elshaug et al[20]	The value of low-value lists	2013	Collection of Studies	Low value care	Target	1	Very Low Quality
				Waste	Target		
Fatovich[21]	Medical reversal: what are you doing wrong for your patient today?	2013	Collection of Studies	Medical reversal	Phenomenon	1	Very Low Quality
Finn and Greenwald[22]	Update in hospital medicine: evidence you should know	2015	Collection of Studies	Recommendations for practice	Consequence	4	Low Quality

Author(s) or Organization	Title	Year	Study Type	Term(s) Used	Relationship to Reversal	AMSTAR Rating Score	AMSTAR Quality Rating
Garner et al[23]	Reducing ineffective practice: challenges in identifying low-value health care using Cochrane Systematic Reviews	2013	Collection of Studies	Low-value health care	Target	2	Very Low Quality
Gnjidic and Elshaug[24]	De-adoption and its 43 related terms: harmonizing low-value care terminology	2015	Collection of Studies	Abandon*	Consequence	2	Very Low Quality
				Contradict	Phenomenon		
				De-adoption	Consequence		
				Decrease use	Consequence		
				Decline in use	Consequence		
				Discontinu*	Consequence		
				Disinvestment	Consequence		
				Ineffective	Target		
				Obsole*	Target		
				Opportunity cost	Phenomenon		
				Overdiagnosis	Phenomenon		
				Overtreatment	Consequence		
				Re-assessment	Consequence		
				Resource re-allocation	Consequence		
				Reversal	Phenomenon		
				Waste	Target		
Haas et al[25]	Breaking up is hard to do: why disinvestment in medical technology is harder than investment	2012	Collection of Studies	Disinvestment	Consequence	1	Very Low Quality
				Improper use	Target		
				Negative list	Target		
				Obsolete / outmoded / abandoned technologies	Target		
				Services not medically necessary	Target		
Hampton[26]	Clinical trial results may lead to changes in cardiovascular care	2014	Collection of Studies	Change in treatment guidelines	Consequence	1	Very Low Quality

Author(s) or Organization	Title	Year	Study Type	Term(s) Used	Relationship to Reversal	AMSTAR Rating Score	AMSTAR Quality Rating
Hanrahan et al[27]	Sacred cow gone to pasture: a systematic evaluation and integration of evidence-based practice	2015	Collection of Studies	Sacred cows	Target	2	Very Low Quality
Ioannidis[28]	Contradicted and initially stronger effects in highly cited clinical research	2005	Systematic Review	Contradicted	Phenomenon	2	Very Low Quality
				Initially stronger effects	Potential Predictor		
Ioannidis[29]	Evolution and translation of research findings: from bench to where?	2006	Collection of Studies	Contradictory results	Phenomenon	1	Very Low Quality
				Proteus phenomenon	Phenomenon		
Ioannidis[30]	Molecular bias	2005	Collection of Studies	Proteus phenomenon	Phenomenon	1	Very Low Quality
Ioannidis[31]	Non-replication and inconsistency in the genome-wide association setting	2007	Systematic Review	Inconsistent results	Potential Predictor	1	Very Low Quality
				Non-replication	Potential Predictor		
Ioannidis and Lau[32]	Evolution of treatment effects over time: empirical insight from recursive cumulative metaanalyses	2001	Recursive Cumulative Meta-Analysis	Uncertainty	Potential Predictor	3	Low Quality
Ioannidis and Panagiotou[33]	Comparison of effect sizes associated with biomarkers reported in highly cited individual articles and in subsequent meta-analyses	2011	Systematic Review	False positive	Potential Predictor	4	Low Quality
				Inflated results	Potential Predictor		
Ioannidis and Trikalinos[34]	Early extreme contradictory estimates may appear in published research: the Proteus phenomenon in molecular genetics research and randomized trials	2005	Systematic Review	Proteus phenomenon	Phenomenon	2	Very Low Quality

Author(s) or Organization	Title	Year	Study Type	Term(s) Used	Relationship to Reversal	AMSTAR Rating Score	AMSTAR Quality Rating
Ioannidis et al[35]	Genetic associations in large versus small studies: an empirical assessment	2003	Systematic Review	Discrepancies of effect over time	Potential Predictor	1	Very Low Quality
Kotzeva et al[36]	Adding value to health care through discontinuation of low-value practices: ESSENCIAL Project in Catalonia	2013	Collection of Studies	Low value practices	Target	1	Very Low Quality
Laiteerapong and Huang[37]	The pace of change in medical practice and health policy: collision or coexistence	2015	Collection of Studies	Medical reversal Change in guideline recommendation	Phenomenon Consequence	1	Very Low Quality
Loder et al[38]	Choosing wisely in headache medicine: the American Headache Society's list of five things physicians and patients should question	2013	Collection of Studies	Overused or misused tests and treatments	Target	2	Very Low Quality
Macleod et al[39]	Biomedical research: increasing value, reducing waste	2014	Collection of Studies	Research Waste	Target	2	Very Low Quality
Makic et al[40]	Evidence-based practice habits: putting more sacred cows out to pasture	2011	Collection of Studies	Sacred cows	Target	1	Very Low Quality
Makic et al[41]	Examining the evidence to guide practice: challenging practice habits	2014	Collection of Studies	Practices not supported by the evidence Sacred cows	Target Target	1	Very Low Quality
Makic et al[42]	Putting evidence into nursing practice: four traditional practices not supported by the evidence	2013	Collection of Studies	Sacred cows	Target	2	Very Low Quality

Author(s) or Organization	Title	Year	Study Type	Term(s) Used	Relationship to Reversal	AMSTAR Rating Score	AMSTAR Quality Rating
Malhotra et al[43]	Choosing wisely in the UK: The Academy of Medical Royal Colleges' initiative to reduce the harms of too much medicine	2015	Collection of Studies	Overtreatment	Consequence	0	Very Low Quality
Mayer and Nachtnebel[44]	Disinvesting from ineffective technologies: lessons learned from current programs	2015	Systematic Review	Disinvestment	Consequence	3	Low Quality
				Ineffective technology	Target		
				Obsolete technology	Target		
McCandless et al[45]	Snake oil version 2	2014	Collection of Studies	Snake oil	Target	2	Very Low Quality
McCandless et al[46]	Snake oil superfoods?	2013	Collection of Studies	Snake oil	Target	2	Very Low Quality
McCandless et al[47]	Snake oil supplements?	2010	Collection of Studies	Snake oil	Target	1	Very Low Quality
Mitera et al[48]	Choosing Wisely Canada cancer list: ten low-value or harmful practices that should be avoided in cancer care	2015	Collection of Studies	Low-value practices	Target	2	Very Low Quality
				Harmful practices	Target		
Morgan et al[49]	Update on medical overuse	2015	Systematic Review	Medical overuse	Consequence	3	Low Quality
				Overdiagnosis	Phenomenon		
				Overtreatment	Consequence		
National Institute for Health and Care Excellence (NICE)[50]	National Institute for Health and Care Excellence (NICE) “Do not do” list	Current	Systematic Review	Do not do recommendations	Target	8	High Quality
National Institute for Health and Care Excellence (NICE)[51]	UK Database of Uncertainties about the Effects of Treatments (UK DUETs)	Current	Overview Systematic Review	Known uncertainty	Potential Predictor	2	Very Low Quality

Author(s) or Organization	Title	Year	Study Type	Term(s) Used	Relationship to Reversal	AMSTAR Rating Score	AMSTAR Quality Rating
Niven et al[52]	Towards understanding the de-adoption of low-value clinical practices: a scoping review	2015	Systematic Review	Abandon*	Consequence	5	Low Quality
				Contradict	Phenomenon		
				Change in practice	Consequence		
				Change in use	Consequence		
				Clinical redesign	Consequence		
				De-adopt*	Consequence		
				Decline in use	Consequence		
				De-commission	Consequence		
				Decrease use	Consequence		
				Defunding	Consequence		
				De-implement*	Consequence		
				De-list	Consequence		
				Disadoption	Consequence		
				Discontinu*	Consequence		
				Disinvest*	Consequence		
				Do not do	Target		
				Drop in use	Consequence		
				Evidence-based reassessment	Consequence		
				Health technology reassessment	Consequence		
				Inappropriate use	Target		
				Ineffective	Target		
				Low value practice / intervention	Target		
				Medical reversal	Phenomenon		
				Misuse	Consequence		
				Obsole*	Target		
				Over use	Consequence		
				Reallocation	Consequence		
				Re-appraisal	Consequence		
				Reassess*	Consequence		

Author(s) or Organization	Title	Year	Study Type	Term(s) Used	Relationship to Reversal	AMSTAR Rating Score	AMSTAR Quality Rating
				Redeploy	Consequence		
				Reduc*	Consequence		
				Refute	Phenomenon		
				Re-invest	Consequence		
				Relinquish*	Consequence		
				Remov*	Consequence		
				Replace	Consequence		
				Re-prioritization	Consequence		
				Resource release	Consequence		
				Reversal	Phenomenon		
				Stop*	Consequence		
				Substitutional re-investment	Consequence		
				Withdraw*	Consequence		
				Withdrawing from a service and redeploying resources	Consequence		
Paprica et al[53]	From talk to action: policy stakeholders, appropriateness, and selective disinvestment	2015	Collection of Studies	De-implementation	Consequence	1	Very Low Quality
				Disinvestment	Consequence		
				Harmful practices	Target		
				Inappropriate care	Target		
				Low-value services	Target		
Polisena et al[54]	Case studies that illustrate disinvestment and resource allocation decision-making processes in health care: a Systematic Review	2013	Systematic Review	Disinvestment	Consequence	6	High Quality
				Ineffective technologies	Target		
				Obsolete technologies	Target		
Prasad[55]	Translation failure and medical reversal: two sides of the same coin	2016	Collection of Studies	Medical reversal	Phenomenon	1	Very Low Quality
				Translation failure	Phenomenon		
Prasad and Cifu[56]	A medical burden of proof: towards a new ethic	2012	Collection of Studies	Medical reversal	Phenomenon	2	Very Low Quality

Author(s) or Organization	Title	Year	Study Type	Term(s) Used	Relationship to Reversal	AMSTAR Rating Score	AMSTAR Quality Rating
Prasad and Cifu[57]	Medical reversal: why we must raise the bar before adopting new technologies	2011	Collection of Studies	Medical reversal	Phenomenon	1	Very Low Quality
Prasad and Cifu[58]	The reversal of cardiology practices: interventions that were tried in vain	2013	Systematic Review	Medical reversal	Phenomenon	2	Very Low Quality
Prasad and Ioannidis[59]	Evidence-based de-implementation for contradicted, unproven, and aspiring healthcare practices	2014	Collection of Studies	Contradicted established medical practices	Phenomenon	1	Very Low Quality
				De-implementation	Consequence		
				Unproven medical practices	Target		
Prasad et al[60]	A decade of reversal: an analysis of 146 contradicted medical practices	2013	Systematic Review	Medical reversal	Phenomenon	3	Low Quality
Prasad et al[61]	Reversals of established medical practices: evidence to abandon ship	2012	Collection of Studies	Medical reversal	Phenomenon	1	Very Low Quality
Prasad et al[62]	The frequency of medical reversal	2011	Systematic Review	Medical reversal	Phenomenon	2	Very Low Quality
Rauen et al[63]	Seven evidence-based practice habits: putting some sacred cows out to pasture	2008	Collection of Studies	Sacred cows	Target	1	Very Low Quality
Ray[64]	Advancing evidence-based practice - a quarterly compilation of research updates most likely to change clinical practice. Fall 2012	2012	Collection of Studies	Research updates most likely to change clinical practice	Target	2	Very Low Quality

Author(s) or Organization	Title	Year	Study Type	Term(s) Used	Relationship to Reversal	AMSTAR Rating Score	AMSTAR Quality Rating
Ray[65]	Advancing evidence-based practice - a quarterly compilation of research updates most likely to change clinical practice. Fall 2013	2013	Collection of Studies	Research updates most likely to change clinical practice	Target	2	Very Low Quality
Ray[66]	Advancing evidence-based practice - a quarterly compilation of research updates most likely to change clinical practice. Spring 2013	2013	Collection of Studies	Research updates most likely to change clinical practice	Target	2	Very Low Quality
Ray[67]	Advancing evidence-based practice - a quarterly compilation of research updates most likely to change clinical practice. Spring 2014	2014	Collection of Studies	Research updates most likely to change clinical practice	Target	2	Very Low Quality
Ray[68]	Advancing evidence-based practice - a quarterly compilation of research updates most likely to change clinical practice. Summer 2013	2013	Collection of Studies	Research updates most likely to change clinical practice	Target	2	Very Low Quality
Ray[69]	Advancing evidence-based practice - a quarterly compilation of research updates most likely to change clinical practice. Summer 2014	2014	Collection of Studies	Research updates most likely to change clinical practice	Target	2	Very Low Quality

Author(s) or Organization	Title	Year	Study Type	Term(s) Used	Relationship to Reversal	AMSTAR Rating Score	AMSTAR Quality Rating
Ray[70]	Advancing evidence-based practice - a quarterly compilation of research updates most likely to change clinical practice. Winter 2012	2012	Collection of Studies	Research updates most likely to change clinical practice	Target	2	Very Low Quality
Ray[71]	Advancing evidence-based practice - a quarterly compilation of research updates most likely to change clinical practice. Winter 2013	2013	Collection of Studies	Research updates most likely to change clinical practice	Target	2	Very Low Quality
Ray[72]	Advancing evidence-based practice - a quarterly compilation of research updates most likely to change clinical practice. Winter 2014	2014	Collection of Studies	Research updates most likely to change clinical practice	Target	2	Very Low Quality
Scott and Elshaug[73]	Foregoing low-value care: how much evidence is needed to change beliefs?	2013	Collection of Studies	Low value care	Target	1	Very Low Quality
Selby et al[74]	Creating a list of low-value health care activities in Swiss primary care	2015	Systematic Review	Low-value health care	Target	1	Very Low Quality
Shojania et al[75]	How quickly do Systematic Reviews go out of date? A survival analysis	2007	Overview Systematic Review	Change in evidence	Phenomenon	3	Low Quality
Singh and Gupta[76]	Impactful clinical trials of 2015: what you need to know	2016	Collection of Studies	Trials likely to change practice	Target	1	Very Low Quality

Author(s) or Organization	Title	Year	Study Type	Term(s) Used	Relationship to Reversal	AMSTAR Rating Score	AMSTAR Quality Rating
Sprenger et al[77]	Quantifying low-value services by using routine data from Austrian primary care	2016	Systematic Review	Ineffective services	Target	3	Low Quality
				Low-value health care / services	Target		
				Unnecessary waste	Target		
Sundsted et al[78]	Update in outpatient general internal medicine: practice-changing evidence published in 2014	2015	Collection of Studies	Practice-changing evidence	Phenomenon	5	Low Quality
Szostek et al[79]	Update in outpatient general internal medicine: practice-changing evidence published in 2015	2016	Collection of Studies	Practice-changing evidence	Phenomenon	6	High Quality
Thorlund et al[80]	Can trial sequential monitoring boundaries reduce spurious inferences from meta-analyses?	2009	Systematic Review	False positive results	Potential Predictor	5	Low Quality
Tricoci et al[81]	Scientific evidence underlying the ACC/AHA clinical practice guidelines	2009	Systematic Review	Class II recommendation	Target	4	Low Quality
				Class III recommendation	Target		
Trikalinos et al[82]	Effect sizes in cumulative meta-analyses of mental health randomized trials evolved over time	2004	Systematic Review	Unfavourable or favourable shifts over time	Phenomenon	1	Very Low Quality
U.S. Preventive Services Task Force (USPSTF)[83]	U.S. Preventive Services Task Force (USPSTF) "Grade 'D' recommendations" for preventive health services	Current	Collection of Studies	Grade D Recommendations	Target	4	Low Quality
				I Statement	Target		

Author(s) or Organization	Title	Year	Study Type	Term(s) Used	Relationship to Reversal	AMSTAR Rating Score	AMSTAR Quality Rating
Venkatesh and Schuur[84]	A "Top Five" list for emergency medicine: a policy and research agenda for stewardship to improve the value of emergency care	2013	Collection of Studies	Low value practice	Target	1	Very Low Quality
Wang et al[85]	Responses of specialist societies to evidence for reversal of practice	2015	Systematic Review	Medical reversal	Phenomenon	3	Low Quality
Wellbery and McAteer[86]	When medicine reverses itself: avoiding practice pitfalls	2013	Collection of Studies	Reversal	Phenomenon	1	Very Low Quality
Wootton et al[87]	Unproven therapies in clinical research and practice: the necessity to change the regulatory paradigm	2013	Collection of Studies	Unproven therapies	Target	1	Very Low Quality

APPENDIX C
AMSTAR EVALUATION FOR 87 INCLUDED ARTICLES

Table 3: AMSTAR Evaluation of included articles

Author(s) or Organization	1. a priori design	2. Study selection and data extraction in duplicate	3. Comprehensive literature search	4. Publication status in inclusion	5. List of included and excluded studies	6. Characteristics of included studies	7. Quality of included studies	8. Appropriate conclusions	9. Appropriate pooling of findings	10. Likelihood of publication bias	11. Conflict of interest	Score
Advancing Medical Professionalism to Improve Health Care Foundation[1]	Unclear	Unclear	Yes	Unclear	No	No	No	No	N/A	N/A	No	1
Australia Comprehensive Management Framework for Medicare Benefits Schedule[2]	Yes	Unclear	Unclear	No	No	Yes	Yes	Yes	N/A	N/A	No	4
Brien et al[3]	Yes	Yes	Yes	Yes	No	Yes	No	No	N/A	N/A	No	5
British Medical Journal[4]	Unclear	Unclear	Unclear	Unclear	No	Yes	No	No	N/A	N/A	No	1
Bryson[5]	No	No	Unclear	No	No	Yes	No	No	N/A	N/A	No	1
Choosing Wisely Canada[6]	Unclear	Unclear	Yes	Unclear	No	No	No	No	N/A	N/A	No	1
Cifu and Prasad[7]	No	Unclear	Unclear	No	No	No	No	No	N/A	N/A	No	0
Davidoff [8]	No	No	No	No	No	Yes	No	No	N/A	N/A	No	1
Doust and Del Mar[9]	Unclear	Unclear	Unclear	Unclear	No	Yes	No	No	N/A	N/A	No	1
Drazer et al[10]	Unclear	Unclear	Unclear	Unclear	No	Yes	No	No	N/A	N/A	No	1
Duckett et al[11]	No	Yes	No	No	No	Yes	No	No	N/A	N/A	No	2
Ebell and Grad[12]	Unclear	Yes	No	Unclear	No	Yes	No	No	N/A	N/A	No	2

Author(s) or Organization	1. a priori design	2. Study selection and data extraction in duplicate	3. Comprehensive literature search	4. Publication status in inclusion	5. List of included and excluded studies	6. Characteristics of included studies	7. Quality of included studies	8. Appropriate conclusions	9. Appropriate pooling of findings	10. Likelihood of publication bias	11. Conflict of interest	Score
Ebell and Grad[13]	Yes	Yes	No	Unclear	No	Yes	No	No	N/A	N/A	No	3
Ebell and Grad[14]	Yes	Yes	No	Unclear	No	Yes	No	No	N/A	N/A	No	3
Ebell and Grad[15]	Yes	Yes	Yes	Unclear	No	Yes	No	No	N/A	N/A	No	4
Ebell and Grad[16]	Yes	Yes	No	Unclear	No	Yes	Yes	Yes	N/A	N/A	No	5
Elshaug et al[17]	Yes	Unclear	Unclear	Unclear	No	Yes	No	No	N/A	N/A	No	2
Elshaug et al[18]	No	Unclear	Unclear	Unclear	No	Yes	No	No	N/A	N/A	No	1
Elshaug et al[19]	Yes	Yes	Yes	Yes	No	Yes	No	No	N/A	N/A	No	5
Elshaug et al[20]	Unclear	Unclear	Unclear	Unclear	No	Yes	No	No	N/A	N/A	No	1
Fatovich[21]	Unclear	Unclear	Unclear	No	No	Yes	No	No	N/A	N/A	No	1
Finn and Greenwald[22]	Unclear	Yes	No	Unclear	No	Yes	Yes	Yes	N/A	N/A	No	4
Garner et al[23]	Unclear	Yes	No	Unclear	No	Yes	No	No	N/A	N/A	No	2
Gnjidic and Elshaug[24]	No	Yes	No	No	No	Yes	No	No	N/A	N/A	No	2
Haas et al[25]	No	Unclear	Unclear	No	No	Yes	No	No	N/A	N/A	No	1
Hampton[26]	No	No	No	No	No	Yes	No	No	N/A	N/A	No	1
Hanrahan et al[27]	Yes	Yes	No	Unclear	No	No	No	No	N/A	N/A	No	2
Ioannidis[28]	No	Yes	Unclear	No	No	Yes	No	No	N/A	N/A	No	2
Ioannidis[29]	Yes	No	Unclear	No	No	No	No	No	N/A	N/A	No	1
Ioannidis[30]	Unclear	No	Unclear	Unclear	No	Yes	No	No	N/A	N/A	No	1
Ioannidis[31]	No	No	No	No	No	Yes	No	No	N/A	N/A	No	1
Ioannidis and Lau[32]	Yes	Unclear	Yes	No	No	Yes	No	No	N/A	N/A	No	3
Ioannidis and Panagiotou[33]	Yes	Yes	Yes	No	No	Yes	No	No	N/A	N/A	No	4
Ioannidis and Trikalinos[34]	Yes	Unclear	Unclear	No	No	Yes	No	No	N/A	N/A	No	2

Author(s) or Organization	1. a priori design	2. Study selection and data extraction in duplicate	3. Comprehensive literature search	4. Publication status in inclusion	5. List of included and excluded studies	6. Characteristics of included studies	7. Quality of included studies	8. Appropriate conclusions	9. Appropriate pooling of findings	10. Likelihood of publication bias	11. Conflict of interest	Score
Ioannidis et al[35]	Unclear	Unclear	No	No	No	Yes	No	No	N/A	N/A	No	1
Kotzeva et al[36]	Yes	Unclear	Unclear	No	No	No	No	No	N/A	N/A	No	1
Laiteerapong and Huang[37]	No	Unclear	Unclear	Unclear	No	Yes	No	No	N/A	N/A	No	1
Loder et al[38]	Yes	Unclear	Unclear	No	No	Yes	No	No	N/A	N/A	No	2
Macleod et al[39]	Yes	Unclear	Unclear	No	No	Yes	No	No	N/A	N/A	No	2
Makic et al[40]	Yes	Unclear	Unclear	Unclear	No	No	No	No	N/A	N/A	No	1
Makic et al[41]	Yes	Unclear	Unclear	No	No	No	No	No	N/A	N/A	No	1
Makic et al[42]	Yes	Unclear	Unclear	No	No	Yes	No	No	N/A	N/A	No	2
Malhotra et al[43]	No	Unclear	Unclear	Unclear	No	No	No	No	N/A	N/A	No	0
Mayer and Nachtnebel[44]	Unclear	Unclear	Yes	Yes	No	Yes	No	No	N/A	N/A	No	3
McCandless et al[45]	Yes	Unclear	Unclear	No	No	Yes	No	No	N/A	N/A	No	2
McCandless et al[46]	Yes	Unclear	Unclear	No	No	Yes	No	No	N/A	N/A	No	2
McCandless et al[47]	No	Unclear	Unclear	No	No	Yes	No	No	N/A	N/A	No	1
Mitera et al[48]	Unclear	Yes	No	Unclear	No	Yes	No	No	N/A	N/A	No	2
Morgan et al[49]	Unclear	Yes	Yes	No	No	Yes	No	No	N/A	N/A	No	3
National Institute for Health and Care Excellence (NICE)[50]	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	N/A	N/A	No	8
National Institute for Health and Care Excellence (NICE)[51]	Yes	Unclear	Unclear	Unclear	No	Yes	No	No	N/A	N/A	No	2
Niven et al[52]	Yes	Yes	Yes	Yes	No	Yes	No	No	N/A	N/A	No	5
Paprica et al[53]	Yes	Unclear	Unclear	No	No	No	No	No	N/A	N/A	No	1
Polisena et al[54]	Unclear	Yes	Yes	Yes	No	Yes	Yes	Yes	N/A	N/A	No	6

Author(s) or Organization	1. a priori design	2. Study selection and data extraction in duplicate	3. Comprehensive literature search	4. Publication status in inclusion	5. List of included and excluded studies	6. Characteristics of included studies	7. Quality of included studies	8. Appropriate conclusions	9. Appropriate pooling of findings	10. Likelihood of publication bias	11. Conflict of interest	Score
Prasad[55]	No	No	No	No	No	Yes	No	No	N/A	N/A	No	1
Prasad and Cifu[56]	Yes	Unclear	Unclear	No	No	Yes	No	No	N/A	N/A	No	2
Prasad and Cifu[57]	Yes	Unclear	Unclear	No	No	No	No	No	N/A	N/A	No	1
Prasad and Cifu[58]	Unclear	Yes	No	No	No	Yes	No	No	N/A	N/A	No	2
Prasad and Ioannidis[59]	No	Unclear	Unclear	No	No	Yes	No	No	N/A	N/A	No	1
Prasad et al[60]	Yes	Yes	No	No	No	Yes	No	No	N/A	N/A	No	3
Prasad et al[61]	No	Unclear	Unclear	Unclear	No	Yes	No	No	N/A	N/A	No	1
Prasad et al[62]	Unclear	Yes	No	No	No	Yes	No	No	N/A	N/A	No	2
Rauen et al[63]	Yes	Unclear	Unclear	No	No	No	No	No	N/A	N/A	No	1
Ray[64]	Yes	Unclear	Unclear	No	No	Yes	No	No	N/A	N/A	No	2
Ray[65]	Yes	Unclear	Unclear	No	No	Yes	No	No	N/A	N/A	No	2
Ray[66]	Yes	Unclear	Unclear	No	No	Yes	No	No	N/A	N/A	No	2
Ray[67]	Yes	Unclear	Unclear	No	No	Yes	No	No	N/A	N/A	No	2
Ray[68]	Yes	Unclear	Unclear	No	No	Yes	No	No	N/A	N/A	No	2
Ray[69]	Yes	Unclear	Unclear	No	No	Yes	No	No	N/A	N/A	No	2
Ray[70]	Yes	Unclear	Unclear	No	No	Yes	No	No	N/A	N/A	No	2
Ray[71]	Yes	Unclear	Unclear	No	No	Yes	No	No	N/A	N/A	No	2
Ray[72]	Yes	Unclear	Unclear	No	No	Yes	No	No	N/A	N/A	No	2
Scott and Elshaug[73]	Unclear	Unclear	Unclear	No	No	Yes	No	No	N/A	N/A	No	1
Selby et al[74]	Unclear	Yes	Unclear	Unclear	No	No	No	No	N/A	N/A	No	1
Shojania et al[75]	Yes	Yes	No	No	No	Yes	No	No	N/A	N/A	No	3
Singh and Gupta[76]	Unclear	Unclear	Unclear	Unclear	No	Yes	No	No	N/A	N/A	No	1
Sprenger et al[77]	Yes	Yes	Unclear	Unclear	No	Yes	No	No	N/A	N/A	No	3
Sundsted et al[78]	Yes	Yes	No	No	No	Yes	Yes	Yes	N/A	N/A	No	5
Szostek et al[79]	Yes	Yes	Yes	Unclear	No	Yes	Yes	Yes	N/A	N/A	No	6

Author(s) or Organization	1. a priori design	2. Study selection and data extraction in duplicate	3. Comprehensive literature search	4. Publication status in inclusion	5. List of included and excluded studies	6. Characteristics of included studies	7. Quality of included studies	8. Appropriate conclusions	9. Appropriate pooling of findings	10. Likelihood of publication bias	11. Conflict of interest	Score
Thorlund et al[80]	Yes	Unclear	Yes	No	No	Yes	Yes	Yes	N/A	N/A	No	5
Tricoci et al[81]	Yes	Unclear	Unclear	No	No	Yes	Yes	Yes	N/A	N/A	No	4
Trikalinos et al[82]	Unclear	Unclear	No	No	No	Yes	No	No	N/A	N/A	No	1
U.S. Preventive Services Task Force (USPSTF)[83]	Yes	Unclear	Unclear	No	No	Yes	Yes	Yes	N/A	N/A	No	4
Venkatesh and Schuur[84]	Unclear	Unclear	Unclear	No	No	Yes	No	No	N/A	N/A	No	1
Wang et al[85]	Unclear	Yes	Unclear	Yes	No	Yes	No	No	N/A	N/A	No	3
Wellbery and McAteer[86]	Unclear	Unclear	Unclear	No	No	Yes	No	No	N/A	N/A	No	1
Wootton et al[87]	Unclear	Unclear	Unclear	No	No	Yes	No	No	N/A	N/A	No	1

APPENDIX D
RATIONALE AND EXAMPLES FOR INCLUSION AND EXCLUSION CRITERIA

CLINICAL PRACTICE, RANDOMIZED CONTROLLED TRIAL, EXISTING PRACTICE

In "A decade of reversal: an analysis of 146 contradicted medical practices," Prasad et al. provided little instruction with regards to the methods used to reach their decisions on aspects of study classification. While they had less restrictive criteria for their review – including all studies that were published as original research articles and all studies that tested clinical practices, regardless of study design and whether they were new or existing – they still sorted their studies into these different categories to inform their descriptive statistics of the 'study' population.

Prasad et al.'s methodology for a study's designation as clinical practice was:

"On the basis of published abstracts, articles were classified as to whether they addressed a clinical practice. Articles addressing a medical practice were defined as any investigation that assesses a screening, stratifying, or diagnostic test, a medication, a procedure or surgery, or any change in health care provision systems. Many research articles concern the novel molecular basis of disease or novel insights in pathophysiology. These articles were excluded. When practice information could not be ascertained by abstract alone, full articles were read."[60]

Prasad et al.'s methodology for classifying study design was:

"Two reviewers (C.T., A.V., M.C., J.R., S.Q., S.J.C., D.B., V.G., or S.S.) and V.P. read articles addressing a medical practice in full. ... Methods were classified as one of the following: randomized controlled trial, prospective controlled (but nonrandomized) intervention study, observational study (prospective or retrospective), case-control study, or other methods."[60]

Prasad et al.'s methodology for determining whether a study examined new or existing practice was:

"Two reviewers (C.T., A.V., M.C., J.R., S.Q., S.J.C., D.B., V.G., or S.S.) and V.P. read articles addressing a medical practice in full. On the ba- sis of the abstract, introduction, and discussion, articles were classified as to whether the practice in question was new or existing."[60]

The methodology does further explain that designation was performed in duplicate and that a third party adjudicated any discrepancies between reviewers. However, as is evident from the above excerpts, beyond providing the categories used, Prasad et al. provide no direction on how they placed studies into those categories. We took the categories that were used by Prasad et al. and created definitions and guidelines

to inform each decision. These guidelines were followed as closely as possible, but often there was no clear distinction between two options and the decision was made at the discretion of the screener.

The first two levels of screening ('clinical practice' and 'study design') were not conducted in duplicate for two reasons. Firstly, the restriction of time would have made screening these stages in duplicate an impossibility, as many articles required careful reading of the full text to inform a decision. Secondly, these two levels had the strongest rationale and guidelines for following, and the distinction between whether a study looked at a clinical practice or not, and between the different study designs, was almost always clear. The third level of screening ('new vs. existing practice') was conducted in duplicate (RQ & DS) to increase the likelihood that we would capture all trials that tested established practices (and consequently lead to reversal or confirmation of evidence). All discrepancies were resolved through discussion (RQ & DS) or mediator (JM). This level was the most difficult to screen as the decision of existing or new practice was made based on the evidence provided in each article, not based on the clinical knowledge of the screeners.

Other researchers or clinicians who assess the same articles may designate them differently across any of the levels, but we are confident in our decisions as they were all made from an independent frame of mind and with all of the information presented by the authors themselves.

Level 1 – Study subject matter

This level of screening was two-tiered, however the inclusion of articles was based only on the first tier. In assessing the subject matter of a study, we designated the direct intervention or thing being studied as well as the primary purpose of that investigation. That is to say, we assessed both whether a study investigated a clinical practice, as well as whether the goal of the study was to assess efficacy with regards to diagnosis, harm, prevention, prognosis, or therapy. While only clinical practices were included in our review, to follow the same methods as Prasad et al., the second-tier of designation was performed to allow for further description of the included population of trials and possible subgroup analyses if it was decided they might be relevant to our outcome of interest.

- **Clinical Practice**: All studies that look at: screening, stratifying, or diagnostic tests; any medical intervention such as a medication, procedure, or surgery; or any other change in health care provision systems that might be tested such as dietary or behavioural interventions and vaccinations.
- **Non-Clinical Practice**: All studies that look at a non-clinical element of medicine including animal studies, studies to elucidate pathophysiolocal pathways, or studies of the molecular basis of disease (e.g. genetic association).
- **Diagnosis**: Examines the ability of a test to identify patients with or without a selected disease or condition; or identifies the frequency of the selected disease or condition. Patients are of two distinct groups, those with the selected disease or

condition, and those without the selected disease or condition.

- **Harm**: Examines the harmful effects of an intervention on measurable outcomes, ideally patient-important outcomes. Patients cannot be randomized to one intervention or another.
- **Prevention**: Examines the ability of an intervention to prevent a selected disease or condition; or the ability of a test to identify apparently healthy patients with a selected disease or condition prior to symptom onset. Patients are apparently healthy people examined before symptom and disease onset for whether or not they develop the selected disease or condition after an intervention; and/or identification of protective or risk factors.
- **Prognosis**: Examines the clinical course of a selected disease or condition that has been treated; or the natural history of a selected disease or condition that has not been treated. Patients are unhealthy people examined after symptom and disease onset for measurable outcomes – ideally patient-important outcomes (death, disease, discomfort, disability, and dissatisfaction) – and/or prognostic factors.
- **Therapy**: Examines the effect of an intervention (medication, procedure, or surgery) on measurable outcomes, ideally patient-important. Patients are unhealthy people with a selected disease or condition after symptom onset

Level 2 – Study design

In general, the different study designs were clearly discernible from author's writing and descriptions of their methods. If the study design was not immediately described in the abstract, then the methods of the full text for the article were examined to reach a conclusion. The different study designs had well defined criteria, which allowed the decision as to a study's designation to be reached quickly and accurately. They are:

- **Randomized controlled trial**: Patients were randomized into two or more groups.
- **Prospective controlled (but nonrandomized) intervention study**: Patients were placed into two or more groups but randomization did not occur.
- **Observational study (prospective or retrospective):** Patients were not placed into groups and randomization did not occur, or all patients received the same intervention.
- **Case-control study**: Patients were selected into two groups based on a certain diagnosis or key attribute
- **Other**: Included other study designs not listed above such as review articles, case series, and case studies that did not fit into any of the above categories

Level 3 – New or existing practice

The designation of whether or not a trial (as all included studies by this tier were randomized controlled trials) investigated a new or an existing practice was difficult. As with the other classifications, it derived from what was presented by the authors in the trial being assessed. However, authors were not always clear in their description of prior patterns of use, and they may have over-exaggerated or under-represented a practice's

prior use to suit their background and rationale for conducting the study. As such, this designation required careful assessment of the trial in question, utilizing clues from the abstract, as well as the full text introduction, discussion, and conclusions. All disagreements at this screening level were resolved through discussion.

To guide inclusion/exclusion at this level, both screeners (RQ & DS) used key words and themes that were associated with trials of new or existing practices. These were primarily used when the authors did not explicitly say whether or not a practice is new or existing. An "existing" practice did not have to be an established standard of care to be included: it had to be described as being in use by physicians. Thus, even practices where there is considerable debate surrounding their use were included as the practices themselves are established enough within the medical community to warrant that debate.

Table 4: Key words and themes that were often found in trials of…

New Clinical Practices	Existing Clinical Practices
• Citation of biological plausibility, animal studies, or lab studies in rationale • Use of pharmacokinetic endpoints as primary outcomes • Description of effect of intervention as "unknown" • Description of intervention use in very different populations (e.g. adults vs. premature infants) • Description of the intervention as untested in the population • Described as Phase I or II trial • Described as Phase III trial, citing only previous Phase I or II trials • Regulatory approval for tested indication after the start of trial recruitment	• Reference to guidelines or recommendations of the practice • Mention of "controversy," "debate" or "uncertainty" within the medical literature • Reference to prior 'epidemiological' or 'observational' studies, which suggests that the practice is seeing use outside of experimental settings • Interventions that are dietary / behavioural / supplemental nutrients or vitamins (as these are often adopted before they have been rigorously tested because they are unlikely to cause harm) • Description of intervention use in similar, but different, populations (e.g. infants vs. premature infants) (required careful consideration of the description of use in both populations) • Described as Phase IV trial • References to older Phase III trials of the same intervention • Regulatory approval for tested indication before the start of trial recruitment • Both intervention and control group are existing therapies or interventions that approach a problem from different positions

While these themes were most often found in trials of the above types of interventions, their presence in an article was not definitive in ascribing the status as new or existing. Rather, they served as clues to help guide the decision and the search for further evidence to inform the decision. Only in cases where the authors failed to provide any more rationale or discussion of the interventions in question, were the above key words and themes used as the basis for finalizing the decision to include or exclude the trial.

The greatest difficulty in this stage of screening arose with interventions that were supplementary or commonly used in other settings. The testing of an intervention in a new population was classified as a "new intervention," and as such, it was often difficult determine when a population had seen use before. This commonly occurred in cancer trials where different combinations of drugs are tested for many different cancers to try and find some degree of efficacy or effectiveness. When an intervention was described as being tested in a different stage of cancer than it is currently used (e.g. an intervention that is currently used for advanced metastatic breast cancer, but being tested for efficacy loco-regionally advanced breast cancer), it was counted as a new population excluded. However, if the stage of cancer that was being tested was similar to one where it already sees use (e.g. an intervention that is currently used for a subcategory of Stage II breast cancer, and being tested in the other subcategory of Stage II breast cancer), then the use in that population was deemed established and the trial was included. The same methodology was used to assess the similarity of other populations and interventions. For example, if an intervention was described as currently existing in a certain population and being tested in a similar population (e.g. childhood vaccination as common practice in a sub-Saharan African country, and being tested in a different sub-Saharan African country where it is not common practice), this was counted as an existing intervention and included.

This was a difficulty in the screening that may be designated differently by other researchers or clinicians assessing the same articles. However, it was unavoidable given our time restriction and lack of clinical expertise. Where Prasad et al. had a team of physicians to reach conclusions regarding the established use or novelty of practices, we did not have the same resources available. This difficulty in assessing new populations for existing practices was the most challenging aspect and led to the greatest number of initial disagreements between screeners. When a practice could not be classified as new or existing, the final decision was to include it (i.e. classify the practice as existing) so as to increase the sensitivity of our screening and capture all potential reversals over the 17-year range.

APPENDIX E
DATA EXTRACTION AND ANALYSIS ELEMENTS

GENERAL STUDY INFORMATION, METHODOLOGY, STUDY RESULTS, STUDY CONCLUSIONS, CONFLICTS OF INTEREST, PICOTS ASSESSMENT, RISK OF BIAS RATINGS, GRADE ASSESSMENT

Table 5: Data extraction methods (options and rationale) for database characteristics

General Study Information	Author(s)	Authors of the study or trial.
	Title	Title of the study or trial.
	DOI	Digital Object Identifier for trial or study. "N/A" if none available.
	Date of Publication	Date the trial or study was published.
	Year of Publication	Year the trial or study was published.
	Protocol Registered	Was the trial or study protocol registered, published, or pre-declared? - Yes - No
	Registration Number	The registration number or location of pre-specified trial or study protocol.
	Year of Registration	Year the trial or study was registered.
	Year Started	Year the trial or study was reported as starting.
	Year of Completion	Year the trial or study was reported as completed.
	Duration between Trial Start and Trial Registration	Calculated difference in years between the year the trial or study started and when it was registered.
	Duration between Trial Registration and Publication	Calculated difference in years between the year the trial or study was registered and when it was published.
	Duration between Trial Completion and Publication	Calculated difference in years between the year the trial or study was reported as completed and when it was published.
	Duration since Publication	Calculated difference in years between the year the trial or study was published and the year of this review.
Methodology	Population	Population in which trial or study is conducted. If two or more trials are reported on in the article, then use the first reported trial or study. If two or more populations are reported, then use the first reported population.

	Protocol Population	Did the population (including inclusion and exclusion criteria) remain unchanged from protocol registration to publication? - Yes - No
	Intervention Group	If two groups are reported as intervention and control / placebo / comparison / currently-used-practice, then use groups as reported. If two groups are reported, but none are labelled as control / placebo / comparison / currently-used-practice / intervention then: - If the two reported groups are one low dose / risk and one high dose / risk, then use the high dose / risk group as the intervention group and the low dose / risk group as the comparison. - If the two reported groups are two interventions and neither is the control / placebo / comparison / currently-used-practice, then use the first reported as the intervention group and the second reported as the comparison group. If more than two groups are reported, and one is control / placebo / comparison / currently-used-practice, then use this as the comparison group, and then: - If the other reported groups are low dose / risk and high dose / risk, then use the low dose / risk as the intervention group. - If the other reported groups are interventions and neither are a control / placebo / comparison / currently-used-practice, then use the first non-control / placebo / comparison / currently-used-practice reported as the intervention group. If more than two groups are reported and none are control / placebo / comparison / currently-used-practice, then use the first mentioned group as the intervention and the second mentioned group as the comparison group. If a factorial design with control / placebo / comparison / currently-used-practice and multiple interventions, then use the double control / placebo / comparison / currently-used-practice as the comparison group and the first-mentioned intervention and control / placebo / comparison / currently-used-practice group as the intervention group. If a factorial design with no control / placebo / comparison / currently-used-practice, then the first mentioned group is the intervention group and the second mentioned group is the comparison group.

	Protocol Intervention Group	Did the intervention group remain unchanged from protocol registration to publication? - Yes - No
	Comparison Group	See group selection for intervention group.
	Protocol Comparison Group	Did the comparison group remain unchanged from protocol registration to publication? - Yes - No
	1º Outcome	If only one primary outcome is reported, then use the sole reported primary outcome. If more than one primary outcome is reported, then use the first primary outcome reported. If no outcomes are reported as the primary outcome, then use the most patient-important outcome as the primary outcome. If primary outcomes are reported for both safety and efficacy, then use the efficacy primary outcome.
	Protocol Primary Outcome	Did the primary outcome remain unchanged from protocol registration to publication? - Yes - No
	Favourable or Unfavourable Primary Outcome	Was the primary outcome favourable (e.g. survival) or unfavourable (e.g. mortality)? - Favourable - Unfavorable
	Unfavourable Primary Outcome	If the primary outcome was unfavourable, then use as reported. If the primary outcome was favourable, then use the complementary unfavourable outcome.
	2º Outcome (Major)	If only one secondary outcome is reported, then use the sole reported secondary outcome. If more than one secondary outcome is reported, then use the most patient-important outcome. i.e. mortality
	Protocol Secondary Outcome	Did the secondary outcome remain unchanged from protocol registration to publication? - Yes - No

	Randomization	Was the trial randomized? - Yes - No
	Duration of Follow-Up	If the duration of follow-up for the selected primary outcome is reported, then use this reported duration. If the duration is not reported, but the mean or median length of follow up is, then use the reported mean or median. If the no duration of follow-up for the selected primary outcome is reported, then use the duration of follow-up for the entire trial or study.
	Protocol Follow-Up	Did the time of follow-up or study or trial duration remain unchanged from protocol registration to publication? - Yes - No
	Sample Size	Total number randomized.
	Required Sample Size	Required sample size calculated prior to trial start.
	Protocol Sample Size	Did the required sample size remain unchanged from pre-specification to publication? - Yes - No
Study results	Loss to Follow-Up Total	Total loss to follow-up in entire trial or study. Also reported as withdrawn, lack of outcome information, missing primary outcome data, or protocol violations. If the loss to follow-up is only reported for each group, this is calculated by summing the loss to follow-up in both the selected intervention and comparison groups. If there are multiple groups, calculate the total loss to follow-up by summing the loss to follow-up from all included groups If loss to follow-up is only reported as a percentage, calculated by multiplying the initial number of included patients by the percentage to receive a whole number. If loss to follow-up is not reported, calculated by the difference between initial number of included patients and the number of observations reported for the primary outcome.

	Loss to Follow-Up in Intervention Group	Loss to follow-up in the selected intervention group for the primary outcome. Also reported as withdrawn, lack of outcome, missing primary outcome data, or protocol violations. If only the total loss to follow-up is reported, then it is assumed loss to follow-up is equal in all groups. This is calculated by dividing the total loss to follow-up by the number of groups. If loss to follow-up for the intervention group is only reported as a percentage, calculated by multiplying the initial number of patients included in the intervention group by the percentage to receive a whole number. If loss to follow-up is not reported, calculated by the difference between initial number of patients included in the intervention group and the number of observations reported in the intervention group for the primary outcome.
	Loss to Follow-Up in Comparison Group	Loss to follow-up in the selected comparison group for the primary outcome. Also reported as withdrawn, lack of outcome information, missing primary outcome data, or protocol violations. If only the total loss to follow-up is reported, then it is assumed loss to follow-up is equal in all groups. This is calculated by dividing the total loss to follow-up by the number of groups. If loss to follow-up for the comparison group is only reported as a percentage, calculated by multiplying the initial number of patients included in the comparison group by the percentage to receive a whole number. If loss to follow-up is not reported, calculated by the difference between initial number of patients included in the comparison group and the number of observations reported in the comparison group for the primary outcome.
	p-Value (1º Outcome)	Reported p-value for selected primary outcome. If no p-value reported, then report as "N/A."
	Significant or Not	Is the p-value significant? - SS (p-value is significant) - NS (p-value is not significant)
	Reported Point Estimate	Reported point estimate of effect for selected primary outcome. If both relative and absolute estimates are available, then use absolute values. If crude / unadjusted and adjusted estimates are available, then use the crude / unadjusted values. If no point estimate is reported (common with continuous outcomes), then report as "Unknown."

		Intent-to-treat analysis point estimates are used over modified-intent-to-treat analysis point estimates which are used over per protocol analysis point estimates
	Reported Confidence Interval (1o Outcome)	Reported confidence interval or standard deviation of selected point estimate of effect of selected primary outcome.
	Type of Outcome	If both continuous and dichotomous outcomes are reported for the selected primary outcome, then the dichotomous outcome is preferred. - Continuous - Dichotomous
	Number of Events in Intervention Group	For dichotomous outcomes, number of events of selected primary outcome in the selected intervention group.
	Number in Intervention Group	Number of population randomized to selected intervention group.
	Intervention Group Rate	If dichotomous outcome, $\hat{p}_1$ = intervention group rate: $\hat{p}_1 = \frac{x_1}{n_1}$ x_1 = number of events in intervention group n_1 = number of patients in intervention group
	Number of Events in Comparison Group	For dichotomous outcomes, number of events of selected primary outcome in the selected comparison group.
	Number in Comparison Group	Number of population randomized to selected comparison group.
	Comparison Group Rate	If dichotomous outcome, $\hat{p}_2$ = comparison group rate: $\hat{p}_2 = \frac{x_2}{n_2}$ x_2 = number of events in comparison group n_2 = number of patients in comparison group
	Absolute Risk Difference Lower 95% Confidence Interval Absolute Risk Difference Upper 95% Confidence Interval	If dichotomous outcome, calculated using Newcombe-Wilson hybrid score confidence intervals: $ARD_{LowerLimit} = (\hat{p}_1 - \hat{p}_2) - z_{\alpha/2}\sqrt{U_2(1-U_2)/n_2 + L_1(1-L_1)/n_1}$ $ARD_{UpperLimit} = (\hat{p}_1 - \hat{p}_2) + z_{\alpha/2}\sqrt{U_1(1-U_1)/n_1 + L_2(1-L_2)/n_2}$ $U = \frac{2n\hat{p} + z^2 + z\sqrt{z^2 + 4n(\hat{p}(1-\hat{p}))}}{2(n+z^2)}$

		$L = \frac{2n\hat{p} + z^2 - z\sqrt{z^2 + 4n(\hat{p}(1-\hat{p}))}}{2(n + z^2)}$ $ARD_{LowerLimit}$ = absolute risk difference lower CI interval $ARD_{UpperLimit}$ = absolute risk difference lower CI interval $\hat{p}_2$ = intervention group rate n_2 = number of patients in intervention group U_2 = U calculated for intervention group L_2 = L calculated for intervention group $\hat{p}_1$ = comparison group rate n_1 = number of patients in comparison group U_1 = U calculated for comparison group L_1 = L calculated for comparison group $z_{\alpha/2}$ = z score for 95% confidence interval, α of 0.05 = 1.96
	Absolute Risk Difference Point Estimate	If dichotomous outcome, ARD = absolute risk difference: $ARD = IGR - CGR$ IGR = intervention group rate CGR = comparison group rate - Absolute Risk Decrease if Positive - Absolute Risk Increase if Negative
	Number Needed to Treat Lower 95% Confidence Interval Number Needed to Treat Upper 95% Confidence Interval	If dichotomous outcome, calculated by: $NNT_{LowerLimit} = \frac{1}{ARD_{UpperLimit}}$ $NNT_{UpperLimit} = \frac{1}{ARD_{LowerLimit}}$ $NNT_{LowerLimit}$ = number needed to treat lower CI interval $NNT_{UpperLimit}$ = number needed to treat upper CI interval $ARD_{UpperLimit}$ = absolute risk difference upper CI interval $ARD_{LowerLimit}$ = absolute risk difference lower CI interval If the 95% confidence intervals for absolute risk difference have opposing signs (one is negative and one is positive), then report as "N/A."
	Number Needed to Treat Point Estimate	If dichotomous outcome, NNT = number needed to treat: $NNT = \frac{1}{ARD}$ ARD = absolute risk difference

		- Number Needed to Benefit if Positive - Number Needed to Harm if Negative
	Total Number of Events	If dichotomous outcome, X = total number of events: $X = x_1 + x_2$ x_1 = number of events in intervention group x_2 = number of events in comparison group
	Relative Risk Difference	If dichotomous outcome, RRR = relative risk difference: $RRR = \frac{ARD}{\hat{p}_2}$ ARD = absolute risk difference $\hat{p}_2$ = comparison group rate - Relative Risk Decrease if Positive - Relative Risk Increase if Negative
	Fragility Index (FI) Reverse Fragility Index (RFI)	If dichotomous outcome with a significant p-value, FI calculated by recalculating the two-sided p-value for Fischer's exact test after adding an event to the group with the fewer reported events while subtracting a non-event from that group. This process continues iteratively until the calculated p-value becomes greater or equal to 0.05. The number reported in this review is the number of added events required to change the p-value from significant to non-significant. (Calculated using: www.fragilityindex.com) If dichotomous outcome with a non-significant p-value, RFI calculated by recalculating the two-sided p-value for Fischer's exact test after subtracting an event from the group with the most reported events while adding a non-event from that group. This process continues iteratively until the calculated p-value becomes less than 0.05. The number reported in this review is the number of subtracted events required to change the p-value from non-significant to significant. (Calculated using the reverse fragility index in the 'fragility index' R package: https://cran.r-project.org/web/packages/fragilityindex/README.html)
	Reported Intervention Group Mean or Median	If continuous outcome, the reported mean or median for the selected primary outcome in the selected intervention group.
	Reported Intervention Group Standard Deviation or Interquartile Range	If continuous outcome, the reported standard deviation or interquartile range for the selected primary outcome in the selected intervention group. If uneven CI around estimate, use the more conservative limit

	Reported Control Group Mean or Median	If continuous outcome, the reported mean or median for the selected primary outcome in the selected comparison group.
	Reported Control Group Standard Deviation or Interquartile Range	If continuous outcome, the reported standard deviation or interquartile range for the selected primary outcome in the selected comparison group. If uneven CI around estimate, use the more conservative limit
	Pooled Standard Deviation	If continuous outcome, s_p = pooled standard deviation: $s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$ n_1 = number in intervention group s_1 = standard deviation for intervention group n_2 = number in comparison group s_2= standard deviation for comparison group
	Standard Effect Size Lower 95% Confidence Interval (continuous) Standard Effect Size Upper 95% Confidence Interval (continuous) Note: Not calculated for dichotomous outcomes	If continuous outcome, calculated by: $LowerLimit = d - z_{\alpha/2}s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$ $UpperLimit = d + z_{\alpha/2}s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$ n_1 = number in intervention group n_2 = number in comparison group d = standard effect size s_p = pooled standard deviation $z_{\alpha/2}$ = z score for 95% confidence interval, α of 0.05 = 1.96
	Standard Effect Size Point Estimate	If continuous: d_c = standard effect size; if dichotomous: d_d = standard es $d_c = \frac{\bar{x}_1 - \bar{x}_2}{s_p}$ $d_d = \frac{ARD_{estimate}}{\sqrt{ARD_{variance}}}$ $ARD_{variance} = \frac{T_e * T_{ne}}{n_T^3} + \frac{C_e * C_{ne}}{n_C^3}$ s_p = pooled standard deviation T_e / C_e = Treated / control events $\bar{x}_1$ = mean for intervention group T_{ne} / C_{ne} = Treated / control non-events $\bar{x}_2$= mean for comparison group n_T / n_C = Treated / control sample
	Adequacy of Power (AP)	If dichotomous outcome, fails to meet AP if (in decreasing order of importance): - Risk Difference 95% CI includes delta in either direction. - No delta is reported - Doesn't meet sample size calculation provided in article.

		If continuous, fails to meet AP if (in decreasing order of importance): - Standard effect size 95% CI includes a standard effect size greater or equal to 0.5 in either direction. - Doesn't meet sample size calculation provided in article
Author's conclusions	End Point Conclusions	- Positive conclusions o If selected intervention group is reported as beneficial / better than the selected comparison group, then report as "Positive." - Negative conclusions or No difference o If selected intervention group is reported as not beneficial / harmful / no different than the selected comparison group, then report as "Negative or No Difference."
	Conclusion	Abstract conclusion / article discussion
	Does the article contradict current medical practice?	Based on abstract conclusion and article introduction, results, and discussion. If new practice versus current practice / placebo beneficial, then yes. If new practice versus current practice / placebo not beneficial / harmful / no different, then no. If current practice versus prior / inferior practice not beneficial / harmful / no different, then yes. If current practice versus prior / inferior practice beneficial, then no. o Yes o No
	How does the article contradict current medical practice?	Abstract background / article introduction / article discussion
	Did the abstract conclusion report the primary outcome?	Based on abstract conclusion. o Yes o No
	Was the abstract conclusion based on subgroup analysis?	Based on abstract conclusion. o Yes o No
	Was the abstract conclusion based on a secondary outcome?	Based on abstract conclusion. o Yes o No
	Was the article withdrawn or retracted?	Was the article withdrawn or retracted? o Yes o No

	If article was withdrawn or retracted, why?		Reason article was withdrawn. Separate retraction article or from Retraction Watch website http://retractionwatch.com/
	Conclusion *based on trial authors' conclusions not systematic review.		- Reversal: current research shows current practice is ineffective or harmful. - Confirmation: current research shows current practice is superior to previous standard of practice/ is effective/ is beneficial.
	Reversal Type		If the conclusion is a reversal, then report if the reversal is due to "Harm outweighs benefits," "Not effective," "Less effective, but still beneficial," or "Beneficial if thought harmful/not-effective/inferior." Otherwise report as "N/A"
	True Reversal		○ Yes: GRADE Rating is High Quality ○ No: GRADE Rating is not High Quality
Conflict(s) of Interest	Conflicts of Interest		If both industry and non-industry reported, then use industry. ○ Industry ○ Non-Industry ○ None Disclosed
	Sources of Funding		Listed sources of funding
PICOTS Assessment	Patient	Description	Sufficient: ○ The 'right' patient population is identified. ○ The patient population is appropriately generalizable or restricted, included or excluded. ○ The 'right' setting is identified. ○ The setting is appropriately generalizable or restricted (multi-centre or single-centre). ○ Appropriately similar to protocol registration. Insufficient: ○ The 'right' patient population is not identified. ○ The patient population is inappropriately generalizable or restricted, included or excluded. ○ The 'right' setting is not identified. ○ The setting is inappropriately generalizable or restricted (multi-centre or single-centre). ○ Inappropriately different from protocol registration.

		Reason	Reasons behind the "Sufficient" or "Insufficient" judgement call using the above justification(s).
	Sample Size	Description	Sufficient: o Actual sample size is greater or equal to required sample size. Insufficient: o Actual sample size is less than required sample size.
		Reason	Reasons behind the "Sufficient" or "Insufficient" judgement call using the above justification(s).
	Intervention	Description	Sufficient: o An appropriate intervention is identified. o The dosage used is specified and scientifically justified. o The frequency of treatment is specified and scientifically justified. o Appropriately similar to protocol registration. Insufficient: o An appropriate intervention is not identified. o The dosage used is not specified or scientifically justified. o The frequency of treatment is not specified or scientifically justified. o Inappropriately different from protocol registration.
		Reason	Reasons behind the "Sufficient" or "Insufficient" judgement call using the above justification(s).
	Comparator	Description	Sufficient: o The comparator is appropriate. o Next best alternative to intervention o Competing alternative to intervention o Standard of care o Gold standard o The placebo is appropriate. o Appears similar to intervention o The dosage used is specified and scientifically justified. o The frequency of treatment is specified and justified. o Appropriately similar to protocol registration. Insufficient: o The comparator is inappropriate. o Inferior to other alternatives / standards of care o Placebo used instead of an existing standard of care o The dosage used is not specified or scientifically justified.

			○ The frequency of treatment is not specified or justified. ○ Inappropriately different from protocol registration.
		Reason	Reasons behind the "Sufficient" or "Insufficient" judgement call using the above justification(s).
	Outcomes	Description	Sufficient: ○ The primary outcome is valid. The secondary outcome(s) are valid. ○ Patient-important, clinically relevant, hard ▪ Affect how a patient functions / feels / survives. ○ Appropriate surrogate (validated) ▪ Surrogate is correlated with the hard outcome of interest. ▪ Surrogate fully captures the net effect of treatment on the hard outcome of interest. ○ Appropriate composite (validated) ▪ Component endpoints of similar importance to patients. ▪ Component endpoints occur with similar frequency. ▪ Component endpoints are likely to have similar relative risk reductions and narrow confidence intervals. ○ If disease-specific mortality is measured, then so is all-cause mortality. ○ The timing / duration of outcome measurement are appropriate. ○ Appropriately similar to protocol registration. Insufficient: ○ The primary outcome is invalid. The secondary outcome(s) are invalid. ○ Not patient-important, or a patient important outcome is missed. ○ Inappropriate surrogate (validated) ▪ Surrogate is not correlated with the hard outcome of interest. ▪ Surrogate does not fully capture the net effect of treatment on the hard outcome of interest. ○ Inappropriate composite (validated) ▪ Component endpoints are not of similar importance to patients. ▪ Component endpoints do not occur with similar frequency. ▪ Component endpoints are not likely to have similar relative risks reductions and narrow confidence intervals. ○ If morality is measured, only cause-specific or *x*-year survival is measured. ○ The timing / duration of outcome measurement are inappropriate. ○ Inappropriately different from protocol registration.

		Reason	Reasons behind the "Sufficient" or "Insufficient" judgement call using the above justification(s).
		Type	○ Hard (Patient-Important): Outcomes that affect how a patient functions / feels / survives by improving a patient's quality of life or increasing length of life. ○ Surrogate: Outcomes that do not affect how a patient functions / feels / survives, but are associated with those outcomes. ○ Composite: A grouping of outcomes with varying importance to the patients.
	Study Design	Description	Sufficient: ○ The study design is appropriate and the best possible scenario to answer the question. ○ Diagnosis: Case-Control Study (All patients receive both the gold standard test and new proposed test regardless of their actual diagnosis). ○ Prognosis: Observational Study ○ Therapy: Randomized Controlled Trial ○ Prevention: Randomized Controlled Trial ○ Harm: Case-Control or Observational Study ○ Appropriately similar to protocol registration Insufficient: ○ The study design is inappropriate or not the best possible scenario to answer the question. ○ Inappropriately different from protocol registration.
		Reason	Reasons behind the "Sufficient" or "Insufficient" judgement call using the above justification(s).
	Study Purpose	Description	Sufficient: ○ Purpose / question is easily detectable and clearly phrased. ○ Should [intervention] be used for [health problem]? ○ Should [intervention] versus [comparison] be used for [health problem]? ○ Should [intervention] be used in [population]? ○ Should [intervention] versus [comparison] be used in [population]? ○ Appropriately similar to protocol registration. Insufficient ○ Purpose / question is not detectable or clearly phrased.

			○ Inappropriately different from protocol registration.
		Reason	Reasons behind the "Sufficient" or "Insufficient" judgement call using the above justification(s).
	Overall		Sufficient: ○ All PICOTS were sufficient. Somewhat Insufficient: ○ One or two PICOTS were insufficient in way(s) that would not likely affect the outcome of the study. Clearly Insufficient: ○ One or more PICOTS were insufficient in way(s) that could likely affect the outcome of the study.
Risk of Bias (ROB) Assessment	Sequence Generation	Description	Method used to generate the allocation sequence is described sufficient detail to allow an assessment of whether it should produce comparable groups. - Definitely Low Risk of Bias ○ The investigators describe a random component in the sequence generation process: random number table, computer random number generator, coin tossing, shuffling cards/envelopes, throwing dice, drawing lots. - Probably Low Risk of Bias ○ Sequence generation process is not described but it is clear that the investigators used a random component in their process. - Probably High Risk of Bias ○ Insufficient information about sequence generation process to permit judgment. ○ Sequence generation process is not described and it is unclear whether the investigators used a random component in their process. - Definitely High Risk of Bias ○ The investigators describe a non-random component in the sequence generation process. ○ Usually, the description involves some systematic, non-random approach: odd/even date of birth, day/date of admission, hospital/clinic record number. ○ Other: judgment of clinician, preference of participant, results of laboratory test/series of tests, availability of intervention.

		Reason	Reasons behind the Risk Level of Bias judgement call using the above justification(s).
	Allocation Concealment	Description	Method used to conceal the allocation sequence is described in sufficient detail to determine whether intervention allocations could have been foreseen in advance or, or during, enrolment. - Definitely Low Risk of Bias o Participants and investigators enrolling participants could not foresee assignments because one of the following, or an equivalent method, was used to conceal allocation: central allocation (telephone/ web-based/ pharmacy-controlled randomization); sequentially numbered drug containers of identical appearance; sequentially numbers, opaque, sealed envelopes. - Probably Low Risk of Bias o Allocation Concealment is not described in complete detail but it is clear that the investigators used a method of concealment. - Probably High Risk of Bias o Insufficient information to permit judgment. o This is usually the case if method of concealment is not described or not described in sufficient detail to allow a definite judgment. - Definitely High Risk of Bias o Participants or investigators enrolling participants could possibly foresee assignments and thus introduce selection bias: open random allocation schedule, assignment envelopes (missing sequential numbers, opaque or sealed), alternation/rotation, date of birth, case record number, any other explicitly unconcealed procedure.
		Reason	Reasons behind the Risk Level of Bias judgement call using the above justification(s).
	Blinding	Description	Described all measures used, if any, to blind study participants, personnel, and outcome assessors from knowledge of which intervention a participant received. Provided any information relating to whether the intended blinding was effective. - Definitely Low Risk of Bias o No blinding, but authors judge the outcome and the outcome measurement are not likely to be influenced by lack of blinding.

			o Blinding of participants and key study personnel; and unlikely that blinding could have been broken. o Either participants or some key study personnel were not blinded, but outcome assessment was blinded and the non-blinding of others is unlikely to introduce bias. - Probably Low Risk of Bias o Blinding is not described in detail but it is clear that appropriate blinding has been used. - Probably High Risk of Bias o Insufficient information to permit judgment. o The study did not address this outcome. - Definitely High Risk of Bias o No blinding or incomplete blinding, and outcome or outcome measure is likely to be influenced by lack of blinding. o Blinding of key study and personnel attempted, but likely that blinding could have been broken. o Either participants or key study personnel were not blinded, and the non-blinding of others is likely to introduce bias.
		Reason	Reasons behind the Risk Level of Bias judgement call using the above justification(s).
	Incomplete Outcome Data	Description	Described the completeness of outcome data for each main outcome, including attrition and exclusions from the analysis. Stated whether attrition and exclusions were reported, the numbers in each intervention group (compared with total randomized participants), reasons for attrition/ exclusions were reported, and any re-inclusions in analyses performed by authors. - Definitely Low Risk of Bias o No missing outcome data. o Reasons for missing outcome data unlikely to be related to true outcome. o Missing outcome data balanced in numbers across intervention groups, with similar reasons for missing data across groups. o Missing data have been imputed using appropriate methods. - Probably Low Risk of Bias

			o For dichotomous outcome data, the proportion of missing outcomes compared with observed event rate not enough to have a clinically relevant impact on the intervention effect estimate. o For continuous outcome data, plausible effect size (difference in means or standardized difference in means) among missing outcomes not enough to have a clinically relevant impact on observed effect size. - Probably High Risk of Bias o Insufficient reporting of attrition/exclusions to permit judgment. o The study did not address this outcome. - Definitely High Risk of Bias o Reason for missing outcome data likely to be related to the true outcome, with either imbalance in numbers or reasons for missing data across intervention groups. o For dichotomous outcome data, the proportion of missing outcomes compared with observed event risk enough to induce clinically relevant bias in intervention effect estimate. o For continuous outcome data, plausible effect side among missing outcomes enough to induce clinically relevant bias in observed effect size. o 'As-treated" analysis done with substantial departure of the intervention received from that assigned at randomization. o Potentially inappropriate application of simple imputation. o Loss to follow-up and failure to adhere to the intention to treat principle when indicated.
		Reason	Reasons behind the Risk Level of Bias judgement call using the above justification(s).
	Selective Outcome	Description	Stated how the possibility of selective outcome reporting was examined by the authors, and what was found. - Definitely Low Risk of Bias o The study protocol is available and all of the study's pre-specified (1º and 2º) outcomes that are of interest in the review have been reported in the pre-specified way. - Probably Low Risk of Bias o The study protocol is not available, but it is clear the published reports include all expected outcomes, including those that were pre-specified.

			- Probably High Risk of Bias
			  o The study protocol is not available, and it is unclear if the published reports include all expected outcomes, including those that were pre-specified.
			  o Insufficient information to permit judgment.
			- Definitely High Risk of Bias
			  o Not all of the study's pre-specified primary outcomes have been reported.
			  o One or more primary outcomes are reported using measurements, analysis methods, or subsets / subgroups of the data that were not pre-specified. Continuous measurements that have been: measured multiple times; transformed from "final scores" to "changes from baseline"; or dichotomized to a cut-off in ways that were not pre-specified.
			  o One or more reported primary outcomes of interest in the review are reported incompletely such that they cannot be entered in a meta-analysis.
			  o The study fails to include results for a key outcome that would be expected to have been reported in such a study.
			  o Reporting some outcomes and not others on the basis of the results. Primary outcomes reported as secondary or secondary outcomes reported as primary.

		Reason	Reasons behind the Risk Level of Bias judgement call using the above justification(s).

	Other Sources of Bias	Description	Stated any important concerns about bias not addressed in other domains in the tool.
			- Definitely Low Risk of Bias
			  o The study appears to be free of other sources of bias.
			- Probably Low Risk of Bias
			  o It is unclear if the study is completely free of other sources of bias but any potential bias is not enough to have a clinically relevant impact.
			- Probably High Risk of Bias
			  o Insufficient information to assess whether an important risk of bias exists.
			  o Insufficient rationale or evidence that an identified problem will induce bias.
			  o Unexplained industry involvement

			- Definitely High Risk of Bias o Potential source of bias related to the specific study design used. o Stopped early due to some data-dependent process (including formal-stopping rule). o Extreme baseline imbalance. o Claimed to have been fraudulent. o Carry-over effects in cross-over trials. o Use of un-validated patient-reported outcomes. o Evidence of potential commercial exploitation (e.g. Industry role in trial design/conduct/analysis)
		Reason	Reasons behind the Risk Level of Bias judgement call using the above justification(s).
	Overall Risk of Bias		- Definitely Low Risk of Bias o All risk of bias items are judged as definitely low risk of bias. o Most risk of bias items are judged as definitely low risk of bias but one or two are probably low risk of bias in way(s) that would not likely affect the outcome of the study. - Probably Low Risk of Bias o Most risk of bias items are judged as definitely low risk of bias or probably low risk of bias but one or two are probably high risk of bias in way(s) that would not likely affect the outcome of the study. - Probably High Risk of Bias o At least one risk of bias item is judged as probably high risk of bias in way(s) that could likely affect the outcome of the study. o One risk of bias item is judged as definitely high risk of bias in a way that would not likely affect the outcome of the study. - Definitely High Risk of Bias o At least one risk of bias item is judged as definitely high risk of bias in way(s) that could likely affect the outcome of the study.
GRADE Assessment	Study Design	Type	Trial: Study design was determined to be a "Randomized Controlled Trial" or "Prospective Controlled Intervention Study"
		Initial Level of Confidence	High: Study type is a "Trial"
	Risk of Bias	Description	Overall ROB Assessment

		Downgrade	Downgrade of 0 o Overall ROB Assessment is judged to be "Definitely Low Risk of Bias" or "Probably Low Risk of Bias" Downgrade of -1 o Overall ROB Assessment is judged to be "Probably High Risk of Bias" Downgrade of -2 o Overall ROB Assessment is judged to be "Definitely High Risk of Bias"
	Directness and Applicability	Description	Overall PICOTS Assessment
		Downgrade	Downgrade of 0 o Overall PICOTS Assessment is judged as "Sufficient" Downgrade of -1 o Overall PICOTS Assessment is judged as "Somewhat Insufficient" Downgrade of -2 o Overall PICOTS Assessment is judged as "Clearly Insufficient"
	Imprecision	Description	Adequacy of Power (AP)
		Downgrade	Downgrade of 0 o Meets mOIS threshold. Downgrade of -1 o Does not meet mOIS threshold.
	Publication Bias - Modified	Description	Presence of Selective Outcome Bias
		Downgrade	Downgrade of 0 o Selective Outcome Bias is undetected and judged as "Definitely Low Risk of Bias" or "Probably Low Risk of Bias" Downgrade of -1 o Selective Outcome Bias is suspected and judged as "Probably High Risk of Bias Downgrade of -2 o Selective Outcome Bias is detected and judged as "Definitely High Risk of Bias"
	Total Downgrades		Calculated sum of all the downgrades.
	Overall Quality of Evidence		Trials begin initially at High Quality and can only be downgraded from there. o Total downgrade of 0: High Quality o Total downgrade of -1: Moderate Quality o Total downgrade of -2: Low Quality o Total downgrade of ≤-3: Very Low Quality

APPENDIX F
STATA DO-FILE 1

SETTING UP THE DATABASE FOR ANALYSES

```
/* Before importing the file, the first row which specifies the headings of
each section of the extraction must be deleted. This will not remove any of
the names of the variables, but will allow STATA to use the 'firstrow' command.
Import the excel file, titled "ThesisAnalyses." Use 'firstrow' to specify that
the first row is the variable names. */

import excel "/Users/Riaz/Desktop/EvidenceReversalDataExtraction.xlsx", sheet("Data
Extraction") firstrow

/* There are 161 variables in the excel file, but we have only 20 potential
predictors that we are interested in and an additional 12 that we are using in
our descriptive statistics. Instead of using 'drop' and specifying all
non-included variables, we will use 'keep' to specify the list of variables
that we want to keep in the dataset. Note that although it will not be used in
the regression, DOI and ID will be kept as identifiers. */

keep ID DOI YearofPublication Registered RegistrationnumberorPreSpecif
Durationbetweenstartandregist Durationbetweentrialcompletion
DurationofFollowUpinWeeks DurationofFollowUpinWeeksI FavourableorUnfavourable
SampleSize RequiredSampleSize LosstoFollowUpproportionof
ImputedLosstoFollowUpProport pValuemainoutcome Alteredpvalue SignificantorNot
MeasureofEffect TypeofOutcome TotalNumberofEventsAdjustment FragilityIndex
StandardEffectSizeAll StandardEffectSizeImputed AP EndPointConclusions
Contradictcurrentmedicalpracti PrimaryOutcomereportedinabstr
Basedonsubgroupanalysis Basedonsecondaryoutcome ReversalType Funding Type
Overall OverallBias QualityofEvidence

/* The data has mostly been imported as string variables and must be converted
appropriately. Numeric variables will be converted using the 'destring'
command, while categorical variables will be converted to numeric using the
'encode' command. All original variables will be kept and the newly generated
"converted" variables will be used for the analyses. */

encode Registered, generate (Registration)
encode RegistrationnumberorPreSpecif, generate (RegistrationNumber)
destring Durationbetweenstartandregist, ignore("N/A") generate (Time_StartandReg)
destring Durationbetweentrialcompletion, ignore("N/A") generate (Time_EndandPub)
encode FavourableorUnfavourable, generate (Unfavourable)
destring DurationofFollowUpinWeeks, ignore("N/A") generate (FollowUpTime)
destring RequiredSampleSize, ignore("N/A") force generate (NRequired) /* 'force'
is required as STATA is recognizing a non-numeric character somewhere in the
```

```
extraction file, other than N/A. However upon close inspection of the data,
only the trials with N/A are converted to missing, so the use of 'force' does
not have any impact on the data-conversion with 'destring' */

destring LosstoFollowUpproportionof, ignore("N/A") generate (LosstoFollowUp)
destring pValuemainoutcome, ignore("N/A") generate (PValue)
encode SignificantorNot, generate (Significant)
encode MeasureofEffect, generate (EffectMeasure)
encode TypeofOutcome, generate (Dichotomous)
destring TotalNumberofEventsAdjustment, ignore("N/A") generate (TotalEvents)
destring FragilityIndex, ignore("N/A") generate (Fragility)
destring StandardEffectSizeAll, ignore("N/A" "Unknown") generate (StandardES)
encode AP, generate (AdequacyofPower)
encode EndPointConclusions, generate (Conclusions)
encode Contradictcurrentmedicalpracti, generate (Reversal)
encode PrimaryOutcomereportedinabstr, generate (AbstractOutcomePrimary)
encode Basedonsubgroupanalysis, generate (AbstractOutcomeSubgroup)
encode Basedonsecondaryoutcome, generate (AbstractOutcomeSecondary)
encode ReversalType, generate (ReasonforReversal)
encode Funding, generate (ConflictsofInterest)
encode Type, generate (OutcomeType)
encode Overall, generate (PICOTS)
encode OverallBias, generate (ROB)
encode QualityofEvidence, generate (GRADE)

/* Several variables were imported, not as string, but as numeric because they
contained only numeric data. These will be renamed to more easily identify
them */

rename YearofPublication YearPublished
rename DurationofFollowUpinWeeksI FollowUpTimeImputed
rename SampleSize NTotal
rename Alteredpvalue PValueImputed
rename ImputedLosstoFollowUpProport LosstoFollowUpImputed
rename StandardEffectSizeImputed StandardESImputed

/* When using the 'encode' command, numeric values are assigned alphabetically.
This means that the ordinal variables will need to be fixed so that the quality
ratings appear in the correct order, with the categories that have the highest
quality having the highest number. PICOT is in the correct order, but for ROB,
"Definitely Low" has a value of 2, "Probably High" has a value of 3, and
"Probably Low" has a value of 4, when they should respectively have values of
4, 2, and 3. This same recoding will be applied to the individual components
of the ROB. Similarly, for GRADE, "High" has a value of 1, "Moderate" has a
value of 3, "Low" has a value of 2, and "Very Low" has a value of 4, when
they should respectively have 4, 3, 2, 1. */
```

```
recode ROB (2 = 4) (3 = 2) (4 = 3) (1 = 1), generate (ROB_Overall)
recode GRADE (1 = 4) (3 = 3) (4 = 1) (2 = 2), generate (GRADE_Overall)

/* The recode command is also necessary for most of our binary covariates, for
which all "No" responses have been given values of 1, and all "Yes" responses
were given values of 2. Similarly, other binary variables will be recoded so
that the '0' corresponds to the state of non-interest as follows: Dichotomous
(1) vs. continuous (0) outcome; Conclusions that are positive (1) vs. negative
or no difference (0); */

recode Registration (1 = 0) (2 = 1), generate (_Registration)
recode RegistrationNumber (97 = 0) (1/600 = 1), generate (RegistrationAvailable)
recode Unfavourable (1 = 0) (2 = 1), generate (_Unfavourable)
recode Significant (1 = .) (2 = 0) (3 = 1), generate (_Significant)
recode Dichotomous (1 = 0) (2 = 1), generate (_Dichotomous)
recode AdequacyofPower (1 = 0) (2 = 1), generate (SufficientAP)
recode Conclusions (1 = 0) (2 = 1), generate (_Conclusions)
recode Reversal (1 = 0) (2 = 1), generate (_Reversal)
recode AbstractOutcomePrimary (1 = 0) (2 = 1), generate (_AbsOutcomePrimary)
recode AbstractOutcomeSubgroup (1 = 0) (2 = 1), generate (_AbsOutcomeSubgroup)
recode AbstractOutcomeSecondary (1 = 0) (2 = 1), generate (_AbsOutcomeSecondary)

/* The database should be set for the analyses at this time. Please see the
do-file in APPENDIX G for the STATA code to conduct all descriptive and
logistic regression analyses. */
```

APPENDIX G
STATA DO-FILE 2

CONDUCTING DESCRIPTIVE AND LOGISTIC REGRESSION ANALYSES

```
/* With the database set up after running the do-file from APPENDIX F, this
do-file conducts all of the analyses presented in the Chapter 5: Results. */

/* First set of commands collects all of the information necessary to present
the descriptive statistics found in Tables 5.2, 5.3, and half of Table 5.1.
The first statistics found in Table 5.1 come from the three tiers of screening
and correspond with the PRISMA flow diagram found in Figure 5.1. */

/* Table 5.1:  */

table Conclusions
table Reversal

/* Table 5.2: Overall population descriptive statistics */

table Registration
table RegistrationAvailable if _Registration == 1
mean Time_StartandReg
table Unfavourable
mean FollowUpTime
mean NTotal
mean NRequired
mean LosstoFollowUp
table Significant
table EffectMeasure
table Dichotomous
table OutcomeType
table AbstractOutcomePrimary
table AbstractOutcomeSubgroup
table AbstractOutcomeSecondary
table ReasonforReversal
table PICOTS
table ROB
table GRADE

/* Table 5.3: Descriptive statistics for reversals and reaffirmations */

sort _Reversal
by _Reversal: table Registration
by _Reversal: table RegistrationAvailable if _Registration==1
by _Reversal: summarize Time_StartandReg
```

```
by _Reversal: table Unfavourable
by _Reversal: summarize FollowUpTime NTotal NRequired LosstoFollowUp
by _Reversal: table Significant
by _Reversal: table EffectMeasure
by _Reversal: table Dichotomous
by _Reversal: table OutcomeType
by _Reversal: table AbstractOutcomePrimary
by _Reversal: table AbstractOutcomeSubgroup
by _Reversal: table AbstractOutcomeSecondary
by _Reversal: table PICOTS
by _Reversal: table ROB
by _Reversal: table GRADE

/* The second set of commands perform all of the regression analyses: first
the relationship that each potential predictor may have with the outcome is
looked at individually, then all covariates are included into an overall-
multivariable logistic regression, and then a backwards-stepwise model is
created from all of the covariates of interest. */

/*Table 5.4: Univariable Logistic Regressions for all potential predictors. Also
included are the 'contrast' commands which test the overall significance of the
individual factor variables. To generate the table of beta-coefficients, found
in APPENDIX H, replace each "logistic" command with "logit." */

logistic _Reversal LosstoFollowUp
logistic _Reversal FollowUpTime
logistic _Reversal PValue
logistic _Reversal NTotal
logistic _Reversal TotalEvents
logistic _Reversal Fragility
logistic _Reversal SufficientmOIS
logistic _Reversal StandardES
logistic _Reversal YearPublished
logistic _Reversal Time_StartandReg
logistic _Reversal Time_EndandPub
logistic _Reversal _Registration
logistic _Reversal _AbsOutcomePrimary
logistic _Reversal _AbsOutcomeSubgroup
logistic _Reversal _AbsOutcomeSecondary
logistic _Reversal i.ConflictsofInterest
testparm i.ConflictsofInterest
logistic _Reversal ib3.OutcomeType
testparm i.OutcomeType
logistic _Reversal i.PICOTS
testparm i.PICOTS
logistic _Reversal i.ROB_Overall
```

```
testparm i. ROB_Overall
logistic _Reversal i.GRADE_Overall
testparm i.GRADE_Overall

/* Conduct univariable logistic regressions for the predictors that have missing
data imputed to determine if the imputation effects the relationship. If not,
then the imputed data are used in place of the missing data. */

logistic _Reversal LosstoFollowUpImputed
logistic _Reversal FollowUpTimeImputed
logistic _Reversal PValueImputed
logistic _Reversal StandardESImputed

/* Check the correlation among all of the non-factor variables to see if any of
them have a high degree of correlation and warrant the removal of one from the
overall model. */

correlate LosstoFollowUpImputed FollowUpTimeImputed PValueImputed NTotal
StandardESImputed YearPublished _AbsOutcomePrimary _AbsOutcomeSubgroup
_AbsOutcomeSecondary

/* Full Multivariable Logistic Regression including all covariates; the 'estat gof'
command tests the goodness-of-fit of the model with Pearson; use of 'group(10)'
performs the Hosmer-Lemeshow goodness-of-fit test. */

logistic _Reversal LosstoFollowUpImputed FollowUpTimeImputed PValueImputed
NTotal StandardESImputed YearPublished _Registration _AbsOutcomePrimary
_AbsOutcomeSubgroup _AbsOutcomeSecondary i.ConflictsofInterest ib3.OutcomeType
i.PICOTS i.ROB_Overall i.GRADE_Overall
estat gof
estat gof, group(10)

/* Conduct the likelihood ratio tests necessary to determine the overall effect
of each factor variable in the model. */

estimates store overall
logistic _Reversal LosstoFollowUpImputed FollowUpTimeImputed PValueImputed
NTotal StandardESImputed YearPublished _Registration _AbsOutcomePrimary
_AbsOutcomeSubgroup _AbsOutcomeSecondary ib3.OutcomeType i.PICOTS
i.ROB_Overall i.GRADE_Overall
lrtest overall
logistic _Reversal LosstoFollowUpImputed FollowUpTimeImputed PValueImputed
NTotal StandardESImputed YearPublished _Registration _AbsOutcomePrimary
_AbsOutcomeSubgroup _AbsOutcomeSecondary i.ConflictsofInterest i.PICOTS
i.ROB_Overall i.GRADE_Overall
lrtest overall
```

```
logistic _Reversal LosstoFollowUpImputed FollowUpTimeImputed PValueImputed
NTotal StandardESImputed YearPublished _Registration _AbsOutcomePrimary
_AbsOutcomeSubgroup _AbsOutcomeSecondary i.ConflictsofInterest ib3.OutcomeType
i.ROB_Overall i.GRADE_Overall
lrtest overall
logistic _Reversal LosstoFollowUpImputed FollowUpTimeImputed PValueImputed
NTotal StandardESImputed YearPublished _Registration _AbsOutcomePrimary
_AbsOutcomeSubgroup _AbsOutcomeSecondary i.ConflictsofInterest ib3.OutcomeType
i.PICOTS i.GRADE_Overall
lrtest overall
logistic _Reversal LosstoFollowUpImputed FollowUpTimeImputed PValueImputed
NTotal StandardESImputed YearPublished _Registration _AbsOutcomePrimary
_AbsOutcomeSubgroup _AbsOutcomeSecondary i.ConflictsofInterest ib3.OutcomeType
i.PICOTS i.ROB_Overall
lrtest overall

/* Backwards-Selection Model building with an exit significance level of 0.5.
The use of 'xi' command is necessary because the stepwise procedure does not
allow for factor variables. 'Xi' creates dummy variables that can then be used
in the overall model estimation procedure. */

xi i.ConflictsofInterest i.OutcomeType i.PICOTS i.ROB_Overall i.GRADE_Overall
stepwise, pr(0.5): logistic _Reversal LosstoFollowUpImputed FollowUpTimeImputed
PValueImputed NTotal StandardESImputed YearPublished _Registration
_AbsOutcomePrimary _AbsOutcomeSubgroup _AbsOutcomeSecondary (_IConflicts_2
_IConflicts_3) (_IPICOTS_2 _IPICOTS_3) (_IROB_Overa_2 _IROB_Overa_3
_IROB_Overa_4) (_IGRADE_Ove_2 _IGRADE_Ove_3 _IGRADE_Ove_4)
estat gof
estat gof, group(10)

/* Conduct the likelihood ratio tests to find the overall significance of the
two factor variables that are included in the model. */

logistic _Reversal LosstoFollowUpImputed PValueImputed StandardESImputed
_AbsOutcomeSubgroup i.PICOTS i.ROB_Overall
estimates store backwards
logistic _Reversal LosstoFollowUpImputed PValueImputed StandardESImputed
_AbsOutcomeSubgroup i.ROB_Overall
lrtest backwards
logistic _Reversal LosstoFollowUpImputed PValueImputed StandardESImputed
_AbsOutcomeSubgroup i.PICOTS
lrtest backwards
```

APPENDIX H
RESULTS

SUPPLEMENTARY TABLES AND FIGURES FOR EXTENDED ANALYSES

Table 6. Comparison of population statistics between reversals and reaffirmations

Descriptive Population Statistic	Reversals (n = 331)	Reaffirmation (n = 280)
Trials registered	293 (88%)	249 (89%)
Protocol / registration accessible	257 (88%)	207 (83%)
Mean time between trial start and registration (years)	1.08 (2.45)	1.35 (2.67)
Trials with a primary outcome oriented around harm	253 (76%)	191 (68%)
Mean duration of follow up (weeks)	111.99 (162.47)	119.27 (169.90)
Mean sample size	3430 (11900.5)	3157 (11154.23)
Mean required sample size (where provided)	2205 (5221.36)	2158 (4586.71)
Mean loss to follow up (% total sample size)	0.062 (0.086)	0.079 (0.103)
Trials with significant primary outcomes (P < 0.05)	114 (34%)	162 (58%)
Trials with a primary outcome measure of effect:		
HR (Hazard Ratio)	99 (30%)	67 (24%)
RR (Relative Risk)	75 (23%)	40 (14%)
AR (Absolute Risk)	54 (16%)	49 (18%)
OR (Odds Ratio)	23 (7%)	26 (9%)
ES / MD (Effect Size / Mean Difference)	33 (10%)	32 (12%)
RRR (Relative Risk Reduction)	2 (1%)	3 (1%)
NNT / NNH (Number Needed to Treat / Harm)	0 (0%)	0 (0%)
N/A (Not Available)	45 (14%)	63 (23%)
Trials with a dichotomous primary outcome	267 (81%)	207 (74%)
Trials with a primary outcome that is:		
Hard (patient-important)	156 (47%)	121 (43%)
Composite	109 (33%)	83 (30%)
Surrogate	66 (20%)	76 (27%)
Trials reporting abstract conclusions based on:		
Primary outcome	282 (85%)	238 (85%)
Secondary outcome	120 (36%)	112 (40%)
Subgroup analyses	20 (6%)	32 (11%)
Trials with overall PICOTS designation:		
Sufficient	130 (39%)	113 (40%)
Somewhat insufficient	162 (49%)	132 (47%)
Clearly insufficient	39 (12%)	35 (13%)
Trials with overall ROB designation:		
Definitely low risk of bias	104 (31%)	61 (22%)
Probably low risk of bias	114 (34%)	98 (35%)
Probably high risk of bias	85 (26%)	82 (29%)
Definitely high risk of bias	28 (9%)	39 (14%)
Trials with overall GRADE level of evidence:		
High	41 (12%)	44 (16%)
Moderate	100 (30%)	69 (25%)
Low	71 (22%)	75 (27%)
Very low	119 (36%)	102 (36%)

Table 7. Correlation between potential predictors (not including factor variables)

	Lossto~d	Follow~d	PValue~d	NTotal	Standa~d	YearPu~d	_Ab~mary	_AbsOu~p	_Ab~dary
LosstoFoll~d	1.0000								
FollowUpTi~d	-0.0172	1.0000							
PValueImpu~d	-0.0682	-0.0026	1.0000						
NTotal	-0.0416	0.2139	-0.0346	1.0000					
StandardES~d	0.0130	0.0088	-0.2372	0.0433	1.0000				
YearPublis~d	0.0354	-0.0924	0.0538	0.0753	0.0298	1.0000			
_AbsOut~mary	-0.0378	0.0166	-0.0736	0.0303	0.0666	0.1213	1.0000		
_AbsOutcom~p	0.0271	0.0745	-0.0329	-0.0105	0.0040	-0.0619	-0.0372	1.0000	
_AbsOut~dary	0.0482	0.0047	-0.0610	0.0014	0.0132	-0.0816	0.0242	0.0635	1.0000

Table 8. Logistic regression beta-coefficient estimates for potential predictors.

Covariate	Logistic Regression Analyses		
	Uni-variable	Overall	Backwards-stepwise
Percent participants lost to follow up (imputed)	-1.91898	-0.99256	-1.33550
Duration of follow up (imputed)	-0.00026	-0.00027	N/A
P-value (imputed)	1.74083	1.50488	1.50628
Sample size	2.08e-06	1.89e-06	N/A
Standardized effect size (imputed)	-0.11555	-0.07462	-0.07196
Total number of events	-0.00011	N/A	N/A
Fragility Index	0.00137	N/A	N/A
Sufficient Adequacy of Power	0.03801	N/A	N/A
Year of publication	0.01094	0.00768	N/A
Years between trial start and trial registration	-0.04203	N/A	N/A
Years between trial completion and publication	0.01348	N/A	N/A
Protocol registered	-0.04088	-0.15311	N/A
Abstract conclusion based on primary outcome	0.01549	0.12919	N/A
Abstract conclusion based on subgroup analyses	-0.69637	-0.59737	-0.62669
Abstract conclusion based on secondary outcome	-0.15890	-0.06959	N/A
Conflicts of interest			N/A
Non-industry vs. Industry	0.09531	-0.14225	
None-reported vs. Industry	0.22699	-0.01820	
Type of outcome			N/A
Hard vs. Surrogate	0.39514	0.38554	
Composite vs. Surrogate	0.41359	0.37397	
Overall PICOTS			
Sufficient vs. Clearly insufficient	0.03193	-0.35935	-0.36662
Somewhat insufficient vs. Clearly insufficient	0.09658	-0.24238	-0.20837
Overall ROB			
Definitely low ROB vs. Definitely high ROB	0.86487	0.86599	0.74019
Probably low ROB vs. Definitely high ROB	0.48259	0.47941	0.41797
Probably high ROB vs. Definitely high ROB	0.36729	0.29949	0.30504
Overall GRADE			N/A
High vs. Very low	0.23725	-0.16315	
Moderate vs. Very low	0.32250	-0.00172	
Low vs. Very low	0.19852	0.03025	

Calculating Odds Ratios for relevant unit differences for continuous covariates

OR = e^(β*difference)
95% CI = e^(β lower limit*difference) to e^(β upper limit*difference)

Univariable analyses:

10% increase in 'Percent of participants lost to follow up' = $e^{(-1.92021*0.1)}$ = 0.83
95% CI = $e^{(-3.71201*0.1)}$ to $e^{(-0.12841*0.1)}$ = 0.69 to 0.99

Imputed 10% increase in 'Percent of participants lost to follow up' = $e^{(-1.91898*0.1)}$ = 0.83
95% CI = $e^{(-3.70893*0.1)}$ to $e^{(-0.12903*0.1)}$ = 0.69 to 0.99

52 week increase in "Duration of follow up' = $e^{(-0.00026*52)}$ = 0.99
95% CI = $e^{(-0.00123*52)}$ to $e^{(0.00070*52)}$ = 0.94 to 1.04

Imputed 52 week increase in "Duration of follow up' = $e^{(-0.00026*52)}$ = 0.99
95% CI = $e^{(-0.00123*52)}$ to $e^{(0.00070*52)}$ = 0.94 to 1.04

0.10 increase in 'P-value' = $e^{(1.67535*0.1)}$ = 1.18
95% CI = $e^{(1.06172*0.1)}$ to $e^{(2.28898*0.1)}$ = 1.11 to 1.26

Imputed 0.10 increase in 'P-value' = $e^{(1.74083*0.1)}$ = 1.19
95% CI = $e^{(1.14766*0.1)}$ to $e^{(2.33400*0.1)}$ = 1.12 to 1.26

100 subject increase in 'Sample size' = $e^{(2.08e-06*100)}$ = 1.00
95% CI = $e^{(-0.00001*100)}$ to $e^{(0.00002*100)}$ = 1.00 to 1.00

50 event increase in 'Total number of events' = $e^{(-0.00011*50)}$ = 1.00
95% CI = $e^{(-0.00029*50)}$ to $e^{(0.00008*50)}$ = 1.00 to 1.00

5 event increase in 'Fragility index' = $e^{(0.00137*5)}$ = 1.01
95% CI = $e^{(-0.01459*5)}$ to $e^{(0.01734*5)}$ = 0.93 to 1.09

5-year increase in 'Year of publication' = $e^{(0.01094*5)}$ = 1.06
95% CI = $e^{(-0.02245*5)}$ to $e^{(0.04433*5)}$ = 0.89 to 1.25

5-year increase in 'Years between start and registration = $e^{(-0.04203*5)}$ = 0.81
95% CI = $e^{(-0.11807*5)}$ to $e^{(0.03401*5)}$ = 0.55 to 1.19

5-year increase in 'Years between end and publication' = $e^{(0.01348*5)}$ = 1.07
95% CI = $e^{(-0.11346*5)}$ to $e^{(0.14043*5)}$ = 0.57 to 2.02

Overall logistic analyses:

10% increase in 'Percent of participants lost to follow up' = $e^{(-0.99255*0.1)} = 0.91$
95% CI = $e^{(-2.85571*0.1)}$ to $e^{(0.87060*0.1)} = 0.75$ to 1.09

52 week increase in "Duration of follow up' = $e^{(-0.00027*52)} = 0.99$
95% CI = $e^{(-0.00136*52)}$ to $e^{(0.00081*52)} = 0.93$ to 1.04

0.10 increase in 'P-value' = $e^{(1.50488*0.1)} = 1.16$
95% CI = $e^{(0.87718*0.1)}$ to $e^{(2.13258*0.1)} = 1.09$ to 1.24

100 subject increase in 'Sample size' = $e^{(1.89e-06*100)} = 1.00$
95% CI = $e^{(-0.00001*100)}$ to $e^{(0.00002*100)} = 1.00$ to 1.00

5-year increase in 'Year of publication' = $e^{(0.00768*5)} = 1.04$
95% CI = $e^{(-0.03373*5)}$ to $e^{(0.04910*5)} = 0.85$ to 1.28

Backwards-Stepwise analyses:

10% increase in 'Percent of participants lost to follow up' = $e^{(-1.33550*0.1)} = 0.88$
95% CI = $e^{(-3.16240*0.1)}$ to $e^{(0.49141*0.1)} = 0.73$ to 1.05

0.10 increase in 'P-value' = $e^{(1.506278*0.1)} = 1.16$
95% CI = $e^{(0.88772*0.1)}$ to $e^{(2.12484*0.1)} = 1.09$ to 1.24

APPENDIX I
A PROPOSED TOOLBOX FOR REVERSAL

PROPOSED METHODS FOR ASSESSING SUFFICIENCY AND STABILITY IN RELATION TO REVERSAL

Sufficiency and Stability for the Identification of Evidence Reversal

Reducing the impact of unnecessary reversals will require the identification of practices that have immature evidence among practices that regularly see use, and those for which the adoption process is underway. The evidence base for questionable existing practices must be assessed to support or contradict continued use, and new practices must provide a matured evidence base before recommending their adoption. To this end, we propose several different methods or tools for describing the sufficiency and stability of evidence among both new and established practices. These include: cumulative meta-analysis, trial sequential analysis / monitoring, Bayesian analysis, value of information analysis, GRADE (Grades of Recommendation Assessment, Development, and Education), and the fragility index.

Table 9. Proposed Tools for determining stability and sufficiency of evidence

Proposed Tool	Description
Cumulative meta-analysis	A series of sequential meta-analyses that shows the cumulative evidence for a research question with each new piece of evidence. Very clear visual indicator of both sufficiency and stability.
Trial sequential analysis / monitoring	A statistical method for assessing the conclusions made by cumulative meta-analyses that accounts for multiple testing by creating monitoring boundaries calculated from an optimal information size based on the assumption that all participants are from a single meta-analysis. Very clear statistical indicator of sufficiency.
Bayesian analysis	Incorporates measures of sufficiency and stability (i.e. invariance) into the calculation of likelihoods based on the prior available information. When evidence is sufficient and stable, inferences are good estimators of the truth. Complex indicator of sufficiency and stability.
Value of information analysis	A model for decision-making that utilizes the number of dependent information sources, precision of those sources, and consequent value of information gained from a source (with greater numbers of sources often being redundant and providing lower value). Complex indicator of sufficiency and stability.
GRADE	"Grades of Recommendation, Assessment, Development, and Evaluation" is a quality assessment tool for rating individual studies or collections of studies using biases and study characteristics. Qualitative indicator of sufficiency and stability.
Fragility index	A measure of the fragility of a study's significance based on how few events (switching from event to non-event) would be required to change a statistically significant result a non-significant result. When applied over several studies in support of a claim, a clear indicator of sufficiency.

Cumulative meta-analysis to identify medical reversal

While improving the development of clinical guidelines and increasing knowledge translation are approaches to reducing the prevalence of medical reversals before they are identified, there remains the issue of identifying practices that should be reversals and are needing de-implementation. There is general agreement in the medical literature that the best means of assessing the efficacy and effectiveness of an intervention or practice is a meta-analysis of RCTs.[81] While there are weaknesses in the process of collecting the evidence for a meta-analysis, it is an established statistical method for integrating the results of multiple studies to determine the overall effect of an intervention.[88,89] Applied to the context of medical reversal, where it is necessary to not only understand the efficacy of a practice but whether enough evidence has accumulated to make a decision, an appropriate technique is cumulative meta-analysis.

Where traditional meta-analysis combines all studies into a single summative estimate, cumulative meta-analysis (CM-A) is a process whereby a series of sequential meta-analyses are performed – one each time a new study is conducted on the topic – thereby generating a running estimate that shows the cumulative strength of the evidence for an effect with each additional study.[90] Cumulative meta-analysis is ideal for determining whether we can trust the evidence for an intervention by allowing for the exploration of two values of research that are not addressed by traditional meta-analysis: sufficiency and stability.[90]

A 1992 paper by Lau et al demonstrates the difference in approach between a traditional meta-analysis and a cumulative meta-analysis of the same database.[91] The authors sought to showcase the technique and the value that it provides to practitioners and policy makers in providing more definitive evidence for an intervention's efficacy that is current with each new piece of evidence. As an example for the technique, the authors examined the use of intravenous streptokinase – compared to placebo or no treatment – as a method of reducing total mortality after MI.[91] The total database included 33 trials conducted between 1959 and 1988, with a total of 36974 patients enrolled and randomized to intervention or control.[91] The authors found that the cumulative evidence showed that intravenous streptokinase provided a statistically significant reduction in mortality for acute MI after only eight trials – and a total of 2432 patients (odds ratio = 0.74, 95% CI: 0.59 to 0.92, P = 0.007).[91]

The difference between these statistical techniques is clear, as are the characteristics of sufficiency and stability. While both the conventional and final cumulative MA found strong support favouring the use of intravenous streptokinase versus placebo or no treatment, in reducing total mortality from acute MI ($Z = -8.16$, $P < 0001$), the results from the individual studies in the conventional meta-analysis appear to jump around and there is no definitive pattern on which to base a judgement before the final estimate.[91] On the other hand, the cumulative meta-analysis depicts a strong trend for significance in favour of streptokinase by the 8th study, and all further studies only serve to narrow the confidence intervals of the estimate.[91] The cumulative MA approach suggests that more than 20 trials were conducted unnecessarily, and upwards of 30000

patients were randomized to not receive life-saving treatment. If researchers in the late 1970s had only looked at the cumulative evidence for the intervention, they may have seen the efficacy that was evident in the cumulative trends but not apparent with the conventional meta-analysis, and streptokinase may have been implemented earlier, with fewer resources and patient lives wasted in the pursuit of unnecessary evidence.

Cumulative meta-analysis is a demonstrably powerful technique for identifying whether the evidence for the efficacy of an intervention is stable and sufficient enough to warrant a reversal. This makes it ideal for the identification of current practices that need to be reversed. However, cumulative MA is not the most ideal tool for identifying medical reversals before implementation because it requires that enough evidence exists to exhibit stability and sufficiency to support a decision. This characteristic of cumulative MA could be considered unethical for a practice that is trending towards ineffectiveness or harm but has not yet reached a point of maturity.

Trial sequential analysis / monitoring

While cumulative meta-analyses provide a clear visual indication of whether or not evidence for a claim is sufficient and stable, the conclusions are at an increased risk of being spuriously significant ($P < 0.05$) as a result of repeated testing for significance as data accumulates.[80,92] Trial sequential analysis is a statistical technique that is applied to cumulative meta-analysis to account for this multiple testing by using monitoring boundaries that are based upon an optimal information size.[80,92]

The "information size" (IS) of a meta-analysis is the anticipated number of subjects (i.e. sample size) that is required to detect a pre-specified intervention effect, based on desired risks of Type I and II Error and the expected heterogeneity among included trials, in an adequately powered trial.[80,92] The information size for a meta-analysis should be the same as expected for a single randomized controlled trial, and any meta-analysis conducted before reaching its IS must be evaluated in a way that accounts for the increased risk of Type I Error (i.e. by calculating and utilizing monitoring boundaries).[80,92]

Trial Sequential Monitoring Boundaries provide the limits for significance of effect in meta-analyses that have sparse data.[80,92] Meta-analyses that meet or exceed their IS are considered to have sufficient evidence to support their conclusions.[80,92]

Bayesian analysis

Bayesian analysis is a method of determining the likelihood of future events occurring based on the information that is currently available.[93] The goals of Bayesian analysis are akin to meta-analysis in that it aims to predict or inform decisions based on what is known. The biggest difference between the two approaches (respectively: Bayesian and Frequentist) is the use of existing evidence and prior beliefs to make inferences about probabilities as opposed to basing probabilities off of average values that are conditional on the null hypothesis.[93]

Bayesian analysis is commonly used in clinical decision making and could possibly be used for the identification of targets for reversal.[93] If the available evidence for a claim is neither sufficient nor stable enough to make an inference of acceptable probability (i.e. the probability of intervention X being an appropriate solution to problem Y is less than [threshold percentage]), then it should be tested with an appropriate RCT and studied until the evidence is mature enough, such that the inference reaches the pre-determined threshold.

Value-of-information analysis

Decision-making is often a complex process and it is generally accepted that the more evidence is available to inform a decision, the better any inferences based on that evidence will be.[94] However, when multiple sources of data are dependent on one another, the redundancy in data actually decreases the expected value of information gained by the multiple sources if they had been independent of one another.[94] Value of information analysis is a Bayesian model for decision-making where the posterior distribution density and likelihood function – and consequently the posterior estimate – are calculated based on the number of information sources, the value estimates, and the dependence of the errors (i.e. precision) of the estimate.[94]

Given the usefulness that value of information analysis has for determining the expected contribution of new information sources to an evidence base, it been suggested as a formal tool for determining the value of proposed randomized trials in moving towards the de-implementation of practices that have been established as "unproven" and are consequently potential targets for reversal.[59] It logically follows that if a proposed trial was determined to not provide any new information of value, then the existing evidence base would be considered sufficient and stable enough to inform a decision.

GRADE

The Grades of Recommendation, Assessment, Development, and Education tool is a measure of study quality.[95] It is largely based on the biases that are present in a study and the methodology employed to study a clinical question – often expressed as the sum of four parts of interest: population, intervention, comparison, and outcome (PICO).[95]

While GRADE can be used to provide a quality score for a single article, it is also frequently used in the generation of clinical guidelines: assimilating multiple studies into a cohesive conclusion.[96] The GRADE tool has eight criteria for rating the quality of evidence – the presence of which can either lower or raise the confidence in study conclusions: risk of bias, inconsistency of results, indirectness of evidence, imprecision, probability of publication bias, magnitude of effect, dose-response curve, and residual confounding supporting conclusions.[97] GRADE is well established as being a valid and reliable tool for evaluating the level and quality of evidence, and it therefore would be appropriate as a measure of both sufficiency and stability to identify an evidence reversal.[40,95]

Fragility Index

The Fragility Index is tool that is used to quantify how fragile the results of a controlled trial are by identifying the change in the number of events required to turn a significant result into a non-significant result.[98] There is also an analogue to the fragility index that provides the fragility of a trial in the opposite direction. The Reverse Fragility Index is the number of subjects in a trial that would be required to experience a non-event to take a conclusion from significant to non-significant. This minimum number of patients who would need to have a change of status from non-event to event, or visa-versa, can be compared between trials, with smaller numbers of events indicating a more fragile finding.[98] In an analysis of 399 RCTs published in high-impact journals, the median Fragility Index was 8, 25% of trials had a Fragility Index of 3 or less, and in 53% of trials, the Fragility Index was lower than the number lost to follow up.[98] If the Fragility Index were applied over multiple studies it could provide a potential measure of the sufficiency of evidence as very low numbers would indicate that not enough evidence has been accumulated to support a claim.

Prioritizing the Identification of Targets for Reversal

Each of the above tools has applicability in the field of evidence reversals in establishing whether or not the evidence has matured enough to declare the practice a reversal. These are necessary because concluding the reversal or reaffirmation of a practice on the basis of a single study or trial is inappropriate. However, in having these tools to find practices that should be de-implemented, there remains the logistic problem of identifying the practices upon which to apply these tools. To this end, Prasad and Ioannidis have proposed seven factors for consideration in assigning priority to the testing of unproven medical practices.[59] We support the use of this framework in future reversal research, in conjunction with our proposed toolbox of reversal, and our proposed framework of reversibility.

1. Priority should be given to test practices for which the current evidence base is weakest
2. Priority should be given to interventions which result in significant net financial burden on health payers
3. Priority should be given to practices that have multiple alternative options, especially if the alternatives are of lower cost or less likely to be overturned because of a separate mechanism of action or stronger supporting evidence
4. Priority should be given to test practices with established harms that confer substantial morbidity
5. Priority should be given to test practices for which the cost of testing is far less than ongoing expenditures of the practice
6. Priority should be given to test practices where negative results may have a large impact
7. Priority should be based on the expected value of information to be gained by funding a particular study, at the proposed size and cost, that may inform the de-implementation of a practice

REFERENCES

1. Advancing Medical Professionalism to Improve Health Care Foundation. The American Board of Internal Medicine (ABIM) Foundation Initiative Choosing Wisely. www.choosingwisely.org/. Accessed July 22, 2014.
2. Australia CMF for MBS. Australian Government Department of Health and Ageing, Medicare "Comprehensive Management Framework" environmental scan (Australia). 2015. www.health.gov.au/internet/main/publishing.nsf/Content/ReviewsCMFM. Accessed July 22, 2014.
3. Brien S, Gheihman G, Tse YK, Brynes M, Harrison S, Dobrow MJ. A scoping review of appropriateness of care research activity in Canada from a health system-level perspective. *Healthc Policy*. 2014;9(4):48-61. doi:10.12927/hcpol.2014.23773.
4. British Medical Journal. BMJ's Too Much Medicine. www.bmj.com/too-much-medicine. Accessed July 22, 2014.
5. Bryson GL. Back to the future: Medical reversals and perioperative medicine. *Can J Anesth*. 2014;61:215-219. doi:10.1007/s12630-013-0103-8.
6. Choosing Wisely Canada. The Canadian Medical Association (CMA) Campaign Choosing Wisely. www.choosingwiselycanada.org/. Accessed July 22, 2014.
7. Cifu AS, Prasad VK. Medical debates and medical reversal. *J Gen Intern Med*. 2015;30(12):1729-1730. doi:10.1007/s11606-015-3481-5.
8. Davidoff F. On the undiffusion of established practices. *JAMA Intern Med*. 2015;175(5):809-811. doi:10.1001/jamainternmed.2015.0167.
9. Doust J, Del Mar C. Why do doctors use treatments that do not work? *BMJ*. 2004;328(February):474-475. doi:10.1136/bmj.328.7438.474.
10. Drazer MW, Salama JK, Hahn OM, Weichselbaum RR, Chmura SJ. Stereotactic body radiotherapy for oligometastatic breast cancer: a new standard of care, or a medical reversal in waiting? *Expert Rev Anticancer Ther*. 2016;16(6):625-632. doi:10.1080/14737140.2016.1178577.
11. Duckett SJ, Breadon P, Romanes D. Identifying and acting on potentially inappropriate care. *Med J Aust*. 2015;203(4):1-6. doi:10.5694/mja15.01241.
12. Ebell MH, Grad R. Top 20 research studies of 2011 for primary care physicians. *Am Fam Physician*. 2012;86(9):835-840.
13. Ebell MH, Grad R. Top 20 research studies of 2012 for primary care physicians. *Am Fam Physician*. 2013;88(6):380-386.
14. Ebell MH, Grad R. Top 20 research studies of 2013 for primary care physicians. *Am Fam Physician*. 2014;90(6):397-402.
15. Ebell MH, Grad R. Top 20 research studies of 2014 for primary care physicians. *Am Fam Physician*. 2015;92(5):377-383.
16. Ebell MH, Grad GR. Top 20 research studies of 2015 for primary care physicians. *Am Fam Physician*. 2016;93(1):756-762.
17. Elshaug AG. Building the evidence base for disinvestment from ineffective health care practices: A case study in obstructive sleep apnoea syndrome. 2007;(October).

18. Elshaug AG, Moss JR, Littlejohns P, Karnon J, Merlin TL, Hiller JE. Identifying existing health care services that do not provide value for money. *Med J Aust*. 2009;190(5):269-273.
19. Elshaug AG, Watt AM, Mundy L, Willis CD. Over 150 potentially low-value health care practices: an Australian study. *Med J Aust*. 2012;197(10):556-560. doi:10.5694/mja13.10080.
20. Elshaug AG, Mcwilliams JM, Landon BE. The value of low-value lists. *JAMA J Am Med Assoc*. 2013;309(8):775-776. doi:10.1001/jama.2013.828.
21. Fatovich DM. Medical reversal: What are you doing wrong for your patient today? *EMA - Emerg Med Australas*. 2013;25:1-3. doi:10.1111/1742-6723.12044.
22. Finn KM, Greenwald JL. Update in hospital medicine: Evidence you should know. *J Hosp Med*. 2015;10(12):817-826. doi:10.1002/jhm.2476.
23. Garner S, Docherty M, Somner J, et al. Reducing ineffective practice: challenges in identifying low-value health care using Cochrane systematic reviews. *J Health Serv Res Policy*. 2013;18(1):6-12. doi:10.1258/jhsrp.2012.012044.
24. Gnjidic D, Elshaug AG. De-adoption and its 43 related terms: harmonizing low-value care terminology. *BMC Med*. 2015;13(1):273. doi:10.1186/s12916-015-0511-4.
25. Haas M, Hall J, Viney R, Gallego G. Breaking up is hard to do: Why disinvestment in medical technology is harder than investment. *Aust Heal Rev*. 2012;36:148-152. doi:10.1071/AH11032.
26. Hampton T. Clinical trial results may lead to changes in cardiovascular care. *JAMA - J Am Med Assoc*. 2014;312(19):1957-1959. doi:10.1001/jama.2014.14319.
27. Hanrahan K, Wagner M, Matthews G, et al. Sacred cow gone to pasture: A systematic evaluation and integration of evidence-based practice. *Worldviews Evidence-Based Nurs*. 2015;12(1):3-11. doi:10.1111/wvn.12072.
28. Ioannidis JPA. Contradicted and initially stronger effects in highly cited clinical research. *JAMA J Am Med Assoc*. 2005;294(2):218-228. doi:10.1001/jama.294.2.218.
29. Ioannidis JP. Evolution and translation of research findings: from bench to where? *PLoS Clin Trials*. 2006;1(7):e36. doi:10.1371/journal.pctr.0010036.
30. Ioannidis JPA. Molecular bias. *Eur J Epidemiol*. 2005;20(9):739-745. doi:10.1007/s10654-005-2028-1.
31. Ioannidis JPA. Non-replication and inconsistency in the genome-wide association setting. *Hum Hered*. 2007;64:203-213. doi:10.1159/000103512.
32. Ioannidis JPA, Lau J. Evolution of treatment effects over time: Empirical insight from recursive cumulative metaanalyses. *Proc Natl Acad Sci United States Am*. 2001;98(3):831-836. doi:10.1126/science.132.3438.1488.
33. Ioannidis JPA, Panagiotou OA. Comparison of effect sizes associated with biomarkers reported in highly cited individual articles and in subsequent meta-analyses. *J Am Med Assoc*. 2011;305(201):2200-2210. doi:10.1001/jama.2011.713.
34. Ioannidis JPA, Trikalinos TA. Early extreme contradictory estimates may appear in published research: The Proteus phenomenon in molecular genetics research and randomized trials. *J Clin Epidemiol*. 2005;58:543-549. doi:10.1016/j.jclinepi.2004.10.019.

35. Ioannidis JPA, Trikalinos TA, Ntzani EE, Contopoulos-Ioannidis DG. Genetic associations in large versus small studies: An empirical assessment. *Lancet*. 2003;361:567-571. doi:10.1016/S0140-6736(03)12516-0.
36. Kotzeva A, Torrente E, Almazán C, et al. Essencial : Adding value to healthcare through discontinuation of low-value practices. In: *2nd Conference of International Society for EBHC 6th International Conference for EBHC Teachers and Developers*. Taormina, Italy; 2013.
37. Laiteerapong N, Huang ES. The pace of change in medical practice and health policy: Collision or coexistence? *J Gen Intern Med*. 2015;30(6):848-852. doi:10.1007/s11606-015-3182-0.
38. Loder E, Weizenbaum E, Frishberg B, Silberstein S. Choosing wisely in headache medicine: The american headache society's list of five things physicians and patients should question. *Headache*. 2013;53:1651-1659. doi:10.1111/head.12233.
39. Macleod MR, Michie S, Roberts I, et al. Biomedical research: Increasing value, reducing waste. *Lancet*. 2014;383:101-104. doi:10.1016/S0140-6736(13)62329-6.
40. Makic MBF, VonRueden KT, Rauen CA, Chadwick J. Evidence-based practice habits: Putting more sacred cows out to pasture. *Crit Care Nurse*. 2011;31(2):38-62. doi:10.4037/ccn2011908.
41. Makic MBF, Rauen C, Watson R, Poteet AW. Examining the evidence to guide practice: Challenging practice habits. *Crit Care Nurse*. 2014;34(2):28-45. doi:10.4037/ccn2014262.
42. Makic MBF, Martin SA, Burns S, Philbrick D, Rauen C. Putting evidence into nursing practice : Four traditional practices not supported by the evidence. *Crit Care Nurse*. 2013;33(2):28-43. doi:10.4037/ccn2013787.
43. Malhotra A, Maughan D, Ansell J, et al. Choosing Wisely in the UK: The Academy of Medical Royal Colleges' initiative to reduce the harms of too much medicine. *BMJ*. 2015;350:h2308. doi:10.1136/bmj.h2308.
44. Mayer J, Nachtnebel A. Disinvesting from ineffective technologies: Lessons learned from current programs. *Int J Technol Assess Health Care*. 2015;31(6):355-362. doi:10.1017/s0266462315000641.
45. McCandeless D. Snake Oil version 2. *Inf is Beautiful*. 2014. www.informationisbeautiful.net/2011/snake-oil-version-2/.
46. McCandless D. Snake Oil Superfoods? *Inf is Beautiful*. 2013. www.informationisbeautiful.net/visualizations/snake-oil-superfoods/.
47. McCandless D. Snake Oil Supplements ? *Inf is Beautiful*. 2010.
48. Mitera G, Earle C, Latosinsky S, et al. Choosing Wisely Canada cancer list : Ten low-value or harmful practices that should be avoided in cancer care. *J Oncol Pract*. 2015;11(3):e296-e303. doi:10.1200/JOP.2015.004325.
49. Morgan D, Wright S, Dhruva S. Update on medical overuse. *JAMA J Am Med Assoc*. 2015;175(1):120-124. doi:10.1001/jamainternmed.2014.5444.
50. National Institute for Health and Care (NICE). National Institute for Health and Care Excellence (NICE) "Do not do" list. 2007. www.nice.org.uk/savingsandproductivity/collection?page=1&pagesize=2000&type=do not do. Accessed July 22, 2014.

51. National Institute for Health and Care (NICE). UK Database of Uncertainties about the Effects of Treatments (UK DUETs). http://www.library.nhs.uk/duets/. Accessed July 22, 2014.
52. Niven DJ, Mrklas KJ, Holodinsky JK, et al. Towards understanding the de-adoption of low-value clinical practices: a scoping review. *BMC Med*. 2015;13:255. doi:10.1186/s12916-015-0488-z.
53. Paprica PA, Culyer AJ, Elshaug AG, Peffer J, Sandoval GA. From talk to action: Policy stakeholders, appropriateness, and selective disinvestment. *Int J Technol Assess Health Care*. 2015;31(4):236-240. doi:10.1017/S0266462315000392.
54. Polisena J, Clifford T, Mitton C, Elshaug AG, Russell E, Skidmore B. Case studies that illustrate disinvestment and resource allocation decision-making processes in health care: A systematic review. *Int J Technol Assess Health Care*. 2013;29(2):174-184. doi:10.1017/S0266462313000068.
55. Prasad V. Translation failure and medical reversal: Two sides to the same coin. *Eur J Cancer*. 2016;52:197-200. doi:10.1016/j.ejca.2015.08.024.
56. Prasad V, Cifu A. A medical burden of proof: Towards a new ethic. *Biosocieties*. 2012;7(1):72-87. doi:10.1057/biosoc.2011.25.
57. Prasad V, Cifu A. Medical reversal: Why we must raise the bar before adopting new technologies. *Yale J Biol Med*. 2011;84:471-478.
58. Prasad V, Cifu A. The reversal of cardiology practices: interventions that were tried in vain. *Cardiovasc Diagn Ther*. 2013;3(4):228-235. doi:10.3978/j.issn.2223-3652.2013.10.05.
59. Prasad V, Ioannidis JP. Evidence-based de-implementation for contradicted, unproven, and aspiring healthcare practices. *Implement Sci*. 2014;9(1):1-5. doi:10.1186/1748-5908-9-1.
60. Prasad V, Vandross A, Toomey C, et al. A decade of reversal: An analysis of 146 contradicted medical practices. *Mayo Clin Proc*. 2013;88(8):790-798. doi:10.1016/j.mayocp.2013.05.012.
61. Prasad V, Cifu A, Ioannidis JPA. Reversals of established medical practices: Evidence to abandon ship. *JAMA J Am Med Assoc*. 2012;307(1):37-38. doi:10.1001/jama.2011.1960.
62. Prasad V, Cifu A. The frequency of medical reversal. *Arch Intern Med*. 2011;171(18):1675-1676. doi:10.1001/archinternmed.2011.295.
63. Rauen CA, Chulay M, Bridges E, Vollman KM, Arbour R. Seven evidence-based practice habits: putting some sacred cows out to pasture. *Crit Care Nurse*. 2008;28(2):98. doi:10.1017/CBO9781107415324.004.
64. Ray IB. Advancing evidence-based practice: A quarterly compilation of research updates most likely to change clinical practice. Fall 2012. *Ochsner J*. 2012;12(3):185-187.
65. Ray IB. Advancing evidence-based practice: A quarterly compilation of research updates most likely to change clinical practice. Fall 2013. *Ochsner J*. 2013;13(3):288-292.
66. Ray IB. Advancing evidence-based practice: A quarterly compilation of research updates most likely to change clinical practice. Spring 2013. *Ochsner J*. 2013;13(1):3-7.

67. Ray IB. Advancing evidence-based practice: A quarterly compilation of research updates most likely to change clinical practice. Spring 2014. *Ochsner J*. 2014;14(1):3-6.
68. Ray IB. Advancing evidence-based practice: A quarterly compilation of research updates most likely to change clinical practice. Summer 2013. *Ochsner J*. 2013;13(2):176-180.
69. Ray IB. Advancing evidence-based practice: A quarterly compilation of research updates most likely to change clinical practice. Summer 2014. *Ochsner J*. 2014;14(2):148-153.
70. Ray IB. Advancing evidence-based practice: A quarterly compilation of research updates most likely to change clinical practice. Winter 2012. *Ochsner J*. 2012;12(4):294-297.
71. Ray IB. Advancing evidence-based practice: A quarterly compilation of research updates most likely to change clinical practice. Winter 2013. *Ochsner J*. 2013;13(4):478-480.
72. Ray IB. Advancing evidence-based practice: A quarterly compilation of research updates most likely to change clinical practice. Winter 2014. *Ochsner J*. 2014;14(4):521-526.
73. Scott IA, Elshaug AG. Foregoing low-value care: How much evidence is needed to change beliefs? *Intern Med J*. 2013;43:107-109. doi:10.1111/imj.12065.
74. Selby K, Gaspoz J-M, Rhodondi N, et al. Creating a list of low-value health care activities in swiss primary care. *JAMA J Am Med Assoc*. 2015;175(4):640-642. doi:10.1001/jamainternmed.2014.8020.
75. Shojania KG, Sampson M, Ansari MT, Ji J, Doucette S, Moher D. How quickly do systematic reviews go out of date ? A survival analysis. *Ann Intern Med*. 2007;147:224-233. doi:10.7326/0003-4819-147-4-200708210-00179.
76. Singh N, Gupta M. Impactful clinical trials of 2015: What clinicians need to know. *Can J Cardiol*. 2016;0(0). doi:10.1016/j.cjca.2013.03.003.
77. Sprenger M, Robausch M, Moser A. Quantifying low-value services by using routine data from Austrian primary care. *Eur J Public Health*. 2016:80. doi:10.1093/eurpub/ckw080.
78. Sundsted KK, Wieland ML, Szostek JH, Post JA, Mauck KF. Update in outpatient general internal medicine : Practice-changing evidence published in 2014. *Am J Med*. 2015;128(10):1065-1069. doi:10.1016/j.amjmed.2015.04.033.
79. Szostek JH, Wieland ML, Post JA, Sundsted KK, Mauck KF. Update in outpatient general internal medicine: Practice-changing evidence published in 2015. *Am J Med*. 2016. doi:10.1016/j.amjmed.2016.03.004.
80. Thorlund K, Devereaux PJ, Wetterslev J, et al. Can trial sequential monitoring boundaries reduce spurious inferences from meta-analyses? *Int J Epidemiol*. 2009;38:276-286. doi:10.1093/ije/dyn179.
81. Tricoci P, Allen JM, Kramer JM, Califf RM, Smith Jr SC. Scientific evidence underlying the ACC / AHA clinical practice guidelines. *J Am Med Assoc*. 2009;301(8):831-841. doi:10.1001/jama.2009.205.
82. Trikalinos TA, Churchill R, Ferri M, et al. Effect sizes in cumulative meta-analyses of mental health randomized trials evolved over time. *J Clin Epidemiol*. 2004;57:1124-1130. doi:10.1016/j.jclinepi.2004.02.018.

83. US Preventative Services Task Force (USPSTF). U.S. Preventive Services Task Force (USPSTF) "Grade 'D' recommendations" for preventive health services. 2016;(May). www.uspreventiveservicestaskforce.org/. Accessed July 22, 2016.
84. Venkatesh AK, Schuur JD. A "top Five" list for emergency medicine: A policy and research agenda for stewardship to improve the value of emergency care. *Am J Emerg Med*. 2013;31:1520-1524. doi:10.1016/j.ajem.2013.07.019.
85. Wang MTM, Gamble G, Grey A. Letter: responses of specialist societies to evidence for reversal of practice. *JAMA Intern Med*. 2015;175(5):845-848. doi:10.1001/jamainternmed.2015.0153.
86. Wellbery C, McAteer R. When medicine reverses itself: avoiding practice pitfalls. *Am Fam Physician*. 2013;88(11):737-738.
87. Wootton SH, Evans PW, Tyson JE. Unproven therapies in clinical research and practice: the necessity to change the regulatory paradigm. *Pediatrics*. 2013;132(4):599-601. doi:10.1542/peds.2013-0778.
88. Marini JJ. Meta-analysis: convenient assumptions and inconvenient truth. *Crit Care Med*. 2008;36(1):328-329. doi:10.1097/01.CCM.0000297959.02114.2B.
89. Hunter A, Williams M. Aggregating evidence about the positive and negative effects of treatments. *Artif Intell Med*. 2012;56:173-190. doi:10.1016/j.artmed.2012.09.004.
90. Muellerleile P, Mullen B. Sufficiency and stability of evidence for public health interventions using cumulative meta-analysis. *Am J Public Health*. 2006;96(3):515-522. doi:10.2105/AJPH.2003.036343.
91. Lau J, Antman EM, Jimenez-Silva J, Kupelnick B, Mosteller F, Chalmers TC. Cumulative meta-analysis of therapeutic trials for myocardial infarction. *N Engl J Med*. 1992;327(4):248-254.
92. Wetterslev J, Thorlund K, Brok J, Gluud C. Trial sequential analysis may establish when firm evidence is reached in cumulative meta-analysis. *J Clin Epidemiol*. 2008;61:64-75. doi:10.1016/j.jclinepi.2007.03.013.
93. Spiegelhalter DJ, Myles JP, Jones DR, Abrams KR. Methods in health service research: An introduction to bayesian methods in health technology assessment. *Br Med J*. 1999;319(7208):508-512. http://www.bmj.com.
94. Clemen RT, Winkler RL. Limits for the precision and value of information from dependent sources. *Oper Res*. 1985;33(2):427-442.
95. Guyatt GH, Oxman AD, Schünemann HJ, Tugwell P, Knottnerus A. GRADE guidelines: a new series of articles in the Journal of Clinical Epidemiology. *J Clin Epidemiol*. 2011;64:380-382. doi:10.1016/j.jclinepi.2010.09.011.
96. Kavanagh BP. The GRADE system for rating clinical guidelines. *PLoS Med*. 2009;6(9):e1000094. doi:10.1371/journal.pmed.1000094.
97. Balshem H, Helfand M, Schunemann HJ, et al. GRADE guidelines: 3. Rating the quality of evidence. *J Clin Epidemiol*. 2011;64:401-406. doi:10.1016/j.jclinepi.2010.07.015.
98. Walsh M, Srinathan SK, McAuley DF, et al. The statistical significance of randomized controlled trial results is frequently fragile: A case for a Fragility Index. *J Clin Epidemiol*. 2014;67:622-628. doi:10.1016/j.jclinepi.2013.10.019.

CURRICULUM VITAE

NAME **QURESHI**, Riaz Gregory

CITIZENSHIP Canadian

EDUCATIONAL BACKGROUND

i) Degrees

2015 - 2017 **Masters of Science, Epidemiology/Biostatistics, Candidate**
Faculty of Epidemiology and Biostatistics
University of Western Ontario, London, ON

2010 - 2015 **Honors Bachelor of Science, Health Studies**
Faculty of Applied Health Sciences
University of Waterloo, Waterloo, ON

CONTINUING EDUCATION

2016 How to Teach Evidence-Based Medicine, International Society for Evidence-Based Healthcare, December 7, 2016

2016 Structural Equation Modeling: An Introduction, Western University, School of Graduate and Post-doctoral Studies, May 10 & 12, 2016

2013 Health Technology Assessment: From Theory to Evidence to Policy. 3-Day Workshop & Forum. Programs for the Assessment of Technology in Health (PATH) Research Institute, Hamilton, ON. June 3 - June 5, 2013

2013 Health Technology Assessment Series: 10 Module Training Program, Drug Information & Research Centre (DIRC), Ontario Pharmacists' Association. PATH Research Institute, McMaster University. June 6 – June 7, 2013

2013 Panel on research ethics. Tri-council Policy Statement: Ethical conduct for research involving humans course on research ethics (TCPS 2: CORE). May 28, 2013 (online course)

2013 McMaster University Chart Review Tutorial: for Researchers Conducting Retrospective Review of Health Records, PATH, McMaster University, Reference # 739094, May 28, 2013

2012 Canada Good Clinical Practices (GCP) Stage 1, Collaborative Institutional Training Initiative (CITI). Reference # 8636239, October 9, 2012

VOLUNTEERING HISTORY

2014 Barista and server, Queen St. Commons Café, Kitchener, ON.

2010 - 2011 Research Intern, Simon Fraser University, Burnaby, BC.

HONOURS & AWARDS

2017 Western Graduate Teaching Assistantship, The Design and Analysis of Clinical Trials, University of Western Ontario

2015 Western Graduate Research Scholarship, University of Western Ontario

2011 - 2014 Dean's Honor List, University of Waterloo

2010 President's Entrance Scholarship, University of Waterloo

EMPLOYMENT HISTORY

2015 – Present: Research Assistant, 'Medical Evidence - Decision Impact - Clinical Integrity' (MEDICI), London Health Sciences Centre. London, ON. (September 2015 – present)

2014 (Jan - Aug): Research Assistant, Programs for the Assessment of Technology in Health (PATH) Research Institute, St. Joseph's Healthcare Hamilton. Hamilton, ON.

2013 (Jun - Aug): Research Assistant, Programs for the Assessment of Technology in Health (PATH) Research Institute, St. Joseph's Healthcare Hamilton. Hamilton, ON.

2012 (Sep - Dec): Research Assistant, Interventional Cardiology Research Group (ICRG), Hamilton Health Sciences. Hamilton, ON.

2012 (Jan - Apr): Research Assistant, Hamilton General Hospital, SMART-AMI Study, Hamilton Health Sciences. Hamilton, ON.

2011 (May - Jul): Research Assistant, British Columbia Center for Excellence in HIV/AIDS: IMPACT-HIV Mathematical Modelling Group, St. Paul's Hospital. Vancouver, BC.

ADMINISTRATIVE RESPONSIBILITIES

2017	Teaching Assistant for graduate and undergraduate course: "Design and Analysis of Clinical Trials" – Taught by Dr. Neil Klar
2016 - Present	Clinical Epidemiology Curriculum Development Committee
2016 - Present	Student Event Coordinator, Department of Epidemiology and Biostatistics
2015 - 2016	Director of Student Rounds, Department of Epidemiology and Biostatistics
2013 - 2015	Research: Teaching Standards and Excellence in Applied Health Sciences

PUBLICATIONS

i) Peer Reviewed Reports

Sutton D, **Qureshi R**, Martin J. Evidence reversal – when new evidence contradicts current claims: a systematic overview review. June 2017. *Accepted for publication in Journal of Clinical Epidemiology)*

Qureshi R, Sutton D, Martin J. A timeline of the evidence-based medicine movement. May 2017. *(Manuscript under review before submission)*

Qureshi R, Sutton D, Martin J. The evolution and use of evidence reversal terminology. May 2017. *(Manuscript under review before submission)*

Qureshi R, Sutton D, Martin J. Approaching evidence reversal and medical reversal – when to say "enough is enough." May 2017. *(Manuscript under review before submission)*

O'Reilly DJ, Bowen JM, Perampaladas K, **Qureshi R**, Xie F, Hughes E. Feasibility of an altruistic sperm donation program in Canada: Results from a population-based model. 2017. Reproductive Health, 14(8). DOI: 10.1186/s12978-016-0275-0

Bowen JM, Campbell K, Sutherland S, Bartlett A, Brooks D, **Qureshi R**, Goldstein R, Gershon AS, Prevost S, Samis L, Kaplan AG, Hopkins RB, MacDougald C, Nunes E, O'Reilly DJ, Goeree R. Pulmonary rehabilitation in Ontario: a cross-sectional survey. Ont Health Technol Assess Ser [Internet]. 2015 March; 15(8): 1-67.

ii) Conference presentations ***note: *italicized name* is the presenter**

Qureshi R, Sutton D, Martin J. (May 30 – June 2, 2017). Evidence Reversal and randomized controlled trials within the NEJM: An analysis of trial characteristics from 2000 to 2016. Conference: Canadian Society for Epidemiology and Biostatistics. Banff, Alberta, Canada. **[Poster Presentation]**

Sutton D, ***Qureshi R***, Martin J. (December 7-9, 2016). Unlocking Evidence Reversal in the literature: a key to terminology. Conference: 5th International Society for Evidence Based Health Care Congress. Kish Island, Iran. [**Oral Presentation**]

Salim M, **Qureshi R**, Sharma K, Anderson K. (March 29, 2016). Recurrence of bipolar disorder during pregnancy: A systematic review. Conference: London Health Research Day. London, ON. [**Poster Presentation**]

Qureshi R, Mielke J. (May 1, 2015). Does ethnicity affect weight gain induced by anti-pychotic medication? A systematic review. Conference: 35th Annual Meeting of the Southern Ontario Neuroscience Association. Hamilton, ON. Conference publication: pp 41; Poster B126. [**Poster Presentation**]

Connolly K, *Dmetrichuk K*, **Qureshi R**, Natarajan M, Schwalm J. (October 27-31, 2012). Barriers to EMS utilization during STEMI. Canadian Journal of Cardiology. Conference: 65th Annual Meeting of the Canadian Cardiovascular Society. Toronto, ON. Conference publication: 28: pp S190. DOI: 10.1016/j.cjca.2012.07.248 [**Poster Presentation**]

U**NPUBLISHED DOCUMENTS / COMMISSIONED REPORTS**

Nam J, Guay N, Cheng H, O'Reilly DJ, **Qureshi R**. The budget impact of the permanent tissue expander-implant for breast reconstruction in post-mastectomy female breast cancer in Ontario: a population-based model. 2013