Medical reversals in primary health care: An exploration of this phenomenon in randomized controlled trials

Christian Ruchon, BSc
Department of Family Medicine
McGill University, Montreal
September 2019

A thesis submitted to McGill University in partial fulfillment of the requirements of the degree of Master of Science in Family Medicine

Abstract

Background

The efficacy of clinical interventions is challenged as new research evidence emerges. The concept of medical reversal (MR) occurs when new evidence determines an established medical practice to be less effective than originally claimed and contributes to practice change. The underlying reasons for MR remain poorly understood in the context of primary healthcare. The purpose of this research is to identify characteristics of randomized controlled trials (RCTs) associated with MR in primary healthcare.

Methodology and methods

A dataset of 960 synopses called Patient-Oriented Evidence that Matters (POEMs) written from 2002-2007 was obtained. These POEMs summarize RCTs selected for their relevance to primary healthcare.

Step 1: From each POEM-RCT, the evidence (E_1) of intervention effect was extracted. Then, evidence about the efficacy of the intervention in each POEM-RCT was extracted from knowledge resources such as DynaMed, in 2019 (E_2). Teams of two physician-raters independently compared the initial (E_1) and updated (E_2) evidence of efficacy and categorized each POEM as (1) reversed or (2) not reversed, in 2019. When an MR was identified, the physician-raters included information about any change in the direction of the effect of the intervention.

Step 2: From each POEM-RCT, factors that may be associated with MR such as sample size, allocation concealment, and level of evidence were extracted.

Standard descriptive statistics and four statistical approaches were used to investigate the relationship between the outcome of interest, MR (Step 1), and factors from Step 2 as independent variables. The results from a multiple logistic regression analysis, a Least Absolute Shrinkage and Selection Operator (LASSO), a classification tree, and a random forest analysis were compared.

Results

Of 408 POEM-RCTs assessed by raters, 34 (8.3% ; 95% CI [6 , 11.4]) were identified as medical reversals in 2019. This represents a rate of reversal of about 2 POEM-RCTs per year. In addition, of the study characteristics investigated, the year, the level of evidence (LOE) and the sample size ratio seem to be the best predictive variables. However, since the number of outcomes of interest is modest (n=34), further confirmatory investigation with a larger sample size is necessary.

Conclusion

We found a relatively low rate of medical reversals in the POEMs database of randomized trials. This helps to support the value of the POEMs database. Further work is required to determine which factors are associated with reversal of POEM-RCTs.

Résumé

Contexte

L'efficacité des interventions cliniques est remise en question à mesure que de nouvelles données de recherche apparaissent. Le concept de renversement de la médecine (RM) se produit lorsque de nouvelles preuves déterminent qu'une pratique médicale établie est moins efficace qu'on ne le prétendait à l'origine et contribue à un changement de pratique. Les raisons sous-jacentes d'une RM restent mal comprises dans le contexte des soins de santé de première ligne. Le but de cette recherche est d'identifier les caractéristiques des essais cliniques randomisés (ECR) associés au phénomène de RM dans les soins de santé primaires.

Méthodologie et méthodes

Un ensemble de 960 *Patient-Oriented Evidence that Matters* (POEM) de 2002 à 2007 sera étudié. Ces POEM, rédigés dans le but d'éduquer les cliniciens, résument les ECR choisis en fonction de leur pertinence pour les soins de santé primaires.

Étape 1 : À partir de chaque POEM-ECR, la preuve (E_1) sera extraite. Ensuite, les preuves de l'efficacité de l'intervention testée dans chaque POEM-ECR seront extraites de ressources de connaissances telles que *DynaMed*, en 2019 (E_2). Des équipes composées de deux médecins évaluateurs indépendants compareront les preuves initiales (E_1) et mises à jour (E2) de l'efficacité et classeront chaque POEM comme (1) renversé ou (2) non renversé, en 2019. Lorsqu'une RM est identifiée, les médecins évaluateurs incluront des informations sur le changement d'orientation de l'effet de l'intervention.

Étape 2 : À partir de chaque POEM-ECR, les facteurs qui peuvent être associés à une RM, comme la taille de l'échantillon, la dissimulation de la répartition des participants à l'essai et le niveau de preuve, seront extraits.

Des statistiques descriptives standard et quatre approches statistiques ont été réalisées pour étudier la relation entre le résultat d'intérêt, la RM (étape 1) comme variable dépendante, et les facteurs de l'étape 2 qui agissaient en tant que variables indépendantes. Les résultats d'une analyse de régression logistique multiple, d'une analyse de régression *LASSO* (*Least Absolute Shrinkage and Selection Operator*), d'un arbre de classification et d'une analyse *random forest* ont été comparés.

Résultats

Sur les 408 POEM-RCT évaluées, 34 (8,3 % ; 95% IC [6 ; 11,4]) ont été identifiées comme des renversements médicaux en 2019. Cela représente un taux de renversement d'environ 2 POEM-ECR par an. De plus, parmi les caractéristiques des ECRs étudiée, l'année, le niveau de preuve (LOE) et le rapport de taille de l'échantillon semblent être les meilleures variables prédictives. Toutefois, comme le nombre de résultats d'intérêt est modeste (n=34), il est nécessaire d'effectuer d'autres études de confirmation avec un échantillon plus important.

Conclusion

Nous avons constaté un taux relativement faible de renversement de la médecine dans la base de données des POEMs d'essais randomisés. Cela aide à soutenir la valeur de la base de données des POEMs. D'autres études sont nécessaires pour déterminer quels facteurs sont associés au renversement des POEM-ECR.

Table of Contents

List of Abbreviations

CMA: Canadian Medical Association

CME: Continuing Medical Education

E_1: Evidence from the bottom line of the POEM

E_2: Evidence from DynaMed

EBM: Evidence-based Medicine

EE+: Essential Evidence Plus

ER: Evidence Reversal

GRADE: Grading of Recommendations Assessment, Development and Evaluation

HRT: Hormone Replacement Therapy

IHC: Informed Health Choices

LOE: Level of Evidence

MR: Medical Reversal

NEJM: New England Journal of Medicine

PCI: Percutaneous Coronary Intervention

POEM: Patient-Oriented Evidence that Matters

POEM-RCT: Patient-Oriented Evidence that Matters on a Randomized Controlled Trial

PSA: Prostate-Specific Antigen

RCT: Randomized Controlled Trial

RoB: Risk of Bias

List of Figures

List of Tables

Acknowledgements

It is with great pleasure that I acknowledge the support and assistance of many individuals, without whom this thesis would not have been possible.

First and foremost, I would like to express my deepest gratitude to my supervisor, Dr. Roland Grad, who has been a role model and a mentor. Your guidance and support have been invaluable. You always led by example, always encouraged me to learn new skills and to collaborate on different projects, and more importantly, to be a diligent and thorough researcher. Over the past two years, you have patiently reviewed my work many times and provided precious feedback, as well as excellent opportunities to present my work. I will always be grateful for this and for the financial support you provided.

I would like to thank my thesis committee members: Dr. Pierre Pluye and Dr. Kristian Filion. For the past two years, Dr. Pierre Pluye has patiently and repeatedly challenged my capacities as a researcher to continually elevate my project. Also, I am grateful for the valuable guidance he offered during my many applications for awards. I am also very grateful for Dr. Kristian Filion's kind contributions with his expertise and time on this topic.

I would like to thank Dr. Tibor Schuster for kindly offering his statistical expertise to support and improve the depth of my analysis. He was a wonderful teacher and took the time to answer many questions.

I also extend my thanks to Dr. Mark Ebell for conceiving the idea for this project.

A special thank you to Dr. Mark Ebell, Dr. Allen Shaughnessy, and Dr. David Slawson for their support and expertise at the early stages that helped shape the foundation and the vision of my project. Furthermore, I am grateful for their help as raters. On that note, a special thank you to all the other raters who kindly participated in the project. Thank you, Dr. Mathieu Rousseau, Dr. Emelie Braschi, Dr. Soumya Sridhar, Dr. Anupriya Grover, Dr. Jennifer Ren-Si Cheung.

I was very fortunate and could not have chosen a better home than the Department of Family Medicine at McGill to pursue graduate studies. I am extremely thankful for Dr. Gillian Bartlett's guidance; she has been an excellent source of inspiration as a researcher, teacher, and mentor. I

would also like to thank Sherrie Child who was a lifesaver many times, tirelessly answering my queries and helping with my applications for awards and scholarships.

I am also very grateful for the multiple scholarships awarded to me by the Department of Family Medicine (travel awards). Moreover, I am grateful for the CIHR Travel Awards that allowed me to present my work at the 46th North American Primary Care Research Group Annual Meeting.

I am also very grateful to the Information Technology Primary Care Research Group (ITPCRG). Your weekly meetings opened the door for me to share ideas and heighten the quality of my project whilst learning from very experienced researchers. Thank you, Quan, Reem, Vera, Josh, Vinita, Michael, Maria, and Sara.

To my FMED family who has been with me every step of the way and with whom I have shared countless experiences that forged my professional and personal life: the dynamic trio (Catherine, Sophia), the Dale Dale crew (Laura, David, and Anish), the *Fwiends* (Alessia and Raquel), Reem, and Vera who provided emotional support whenever I needed it and who let me bounce off all my ideas on them before directing me to the best solution. Mary, Sarah, Erica, Juan, Lashanda, Stephanie, Sarah, Nadia, Yvan, and Tamara for fostering a stimulating environment in which to thrive.

I would also like to recognize the help from my friends who gave me great advice on writing my thesis and on remaining healthy. Thank you, the Broadway crew (Diego, Marie, and Juliette), Sabrina, and Christina.

Thank you to my best friends who cheerfully put up with my complaining and my erratic availability and loved me anyway. Thank you, Rockey, Félix, Charles, and Sarah.

Last but definitely not least, I am grateful for my parents: Andrea and Jean-Pierre. I am and will forever be indebted for everything you have done for me and for the continuous support you have offered in all spheres of my life, including my academic endeavors. I can never thank you enough. Thank you, Hugo and Julien, for being the best siblings I could have wished for. You unknowingly drive me to keep pushing for the better.

Thank you.

Preface

This research stemmed from an idea offered by Dr. Mark Ebell.

The following thesis has the format of a traditional thesis.

I. Introduction

The evidence-based medicine (EBM) model is widely adopted in teaching about evidence-based clinical practice. This model consists of three major domains - clinician expertise, patient preference, and the use of high-quality, updated evidence. These domains interact to optimize clinical decision-making at the point of care (1). However, at the point of care, the implementation of clinical evidence in decision-making remains particularly challenging. The latter remains a difficult endeavor for physicians at the point of care despite the advancements of information technology (2).

It is said the volume of studies published has reached 75 trials and 11 systematic reviews per day (3). Indeed, biomedical research is so dynamic that it is unreasonable to expect a physician to read primary research.

In addition, there are many concerns about the quality of the evidence that is readily available, especially in terms of its trustworthiness. Indeed, many biases can distort the conclusions of scientific papers. This has led some to describe the EBM model as being hijacked into an evidence-*biased* model (4, 5). In fact, most large clinical trials are now industry funded with outcomes that tend to favor their treatments (6, 7). This is alarming considering that physicians are expected to be able to use outcomes from these influential trials to inform individual treatment decisions (8). In other work, researchers have described the current state of the quality of clinical evidence as a "Medical Misinformation Mess", where "much of published medical research is not reliable or is of uncertain reliability, offers no benefit to patients, or is not useful to decision makers" (9). If physicians had the skills to evaluate study quality and applicability, they could reasonably choose whether to integrate the findings of research into their practice. However, this does not seem to be the case (10). The inability to assess which study results should be adopted into practice is particularly important when a clinical intervention or practice is challenged as new research evidence emerges.

Contradictions in the evidence have given rise to the phenomenon of medical reversal (MR) (11). The concept of MR occurs when new evidence determines an established medical practice to be less effective or more harmful than originally claimed and contributes to practice change (12). An example of a reversal is the prescribing of hormone replacement therapy (HRT) to otherwise healthy postmenopausal women. Based on the findings of observational studies,

physicians initially recommended this intervention to decrease cardiovascular disease in the 1990s. Subsequently, a large and well conducted randomized controlled trial (RCT) found HRT to increase cardiovascular events and offer no mortality benefit (13, 14).

The reversal of a clinical practice is not uncommon. Indeed, about 40% of original articles testing standard of care and published in the *New England Journal of Medicine* (NEJM) over a decade contradicted an accepted practice (15). As this estimate of reversal is from a single research study, further studies of this phenomenon would seem to be worthwhile.

The reversal of a medical practice can reveal the harms associated with that practice (11). First, there is the direct harm from an intervention that has been shown to be ineffective or harmful. Second, there is an unnecessary cost of the intervention to the parties involved, whether it is the patients or the taxpayers who fund the healthcare system. Third, patients may lose faith in the health care system when treatments are revealed to be ineffective. This loss of trust can negatively impact the quality of the relationship between doctors and their patients. Finally, once new evidence is published and ultimately reaches clinicians, this new and contradictory evidence will be received with varying degrees of resistance to change. Indeed, any study that contradicts an aspect of a physician's practice will face some resistance before its application in practice. The duration of this resistance can range from days to years. During this time, we can assume that ineffective or harmful treatments will still be used in practice, further causing the harms mentioned above. Nevertheless, [the harms of a medical reversal] such harms are potentially avoidable with a more rigorous process for adoption of interventions into clinical practice (16, 17). By way of contrast, some have also been critical of physicians for being too slow to adopt a proven test or treatment (18).

The phenomenon of MR has been associated with surrogate end-points, comparisons with the wrong controls, overconfidence in pathophysiological reasoning, and extrapolation of study results to an age group not yet studied (11, 19). That being said, the underlying reasons for MR remain poorly understood in the context of primary healthcare. Furthermore, research on this topic has not specifically investigated the phenomenon of MR in RCTs applicable to primary healthcare. Therefore, the objective of this research is to explore the phenomenon of MR in primary healthcare and to identify characteristics of RCTs associated with MR in this context.

An interesting hypothesis that could be explored in future research is as follows: A better understanding of the characteristics of RCTs at higher risk of reversal is of value. Knowing the

factors associated with MR, health professionals would avoid prematurely recommending ineffective tests or treatments based on primary studies at high risk of reversal.

The proposed study falls under a meta-epidemiological methodology (20), which adopts "a systematic review or meta-analysis approach to examine the impact of certain characteristics of clinical studies on the observed effect and provide empirical evidence for hypothesized associations" (21). Indeed, I will examine whether characteristics of randomized controlled trials, as the unit of analysis, are associated with the outcome of interest; medical reversal.

The following meta-epidemiological study is postpositivist as the knowledge produced by RCTs is both circumstantial and subject to reconsideration. Such a perspective differs from the positivist view of evidence in which knowledge is considered as stable and secure through time (22). Before starting my research, I understand and accept that scientific evidence is imperfect and fallible for many reasons: biases, questionable methods, and conflicting interests to name a few. In that sense, treatment claims are always circumstantial to a particular setting. The underlying assumption from this research study is that physicians use a post-positivistic perspective to keep updated with new evidence, whether a new area of research is described or the evidence concerns an already existing conjecture found in the literature.

This research is arguably a primary study. As a reminder, research can be categorized at three levels depending on the investigation (23). Level 1 concerns empirical studies where the primary data is collected from fieldwork or lab work (24). Level 2 concerns systematic reviews where the findings are themselves extracted from the empirical studies included in the review (25). Finally, level 3 concerns reviews of reviews, where the findings arise from the reviews included in the study (25). Although the unit of analysis of my research is the RCT, which are empirical studies, I would argue my research is a primary study for three reasons: 1- My research will examine the relationship between variables and some outcome, here the characteristics of trials and MR; 2- The data collection required fieldwork where the outcome of interest was identified and the characteristics of trials were extracted; 3- Data analysis was necessary to make sense and interpret the findings from the field work. In a way, each trial can be considered as a participant in my study.

In this thesis, I use “I” and “we” interchangeably because I was not always working on my own. It is sometimes difficult to tease out what was done by myself alone from what was done with the help of my supervisor.

II. Literature Review

2.1 Medical reversal – when new evidence is inconsistent with established practice

A medical reversal (MR) occurs when new evidence shows an established medical practice to be less effective or more harmful than originally claimed and contributes to practice change (11). This new evidence has to be of a superior quality, due to better study design and increased statistical power for example. This definition of MR comes from the work of Vinay Prasad and colleagues. MRs are not medical interventions that were replaced by better ones, but rather an intervention that was reversed because it was found to be harmful or no better than doing nothing. An example of a reversal is the prescribing of hormone replacement therapy to otherwise healthy postmenopausal women. Based on the findings of observational studies, physicians initially recommended this intervention to decrease cardiovascular disease, but this was later found to be ineffective (13, 14).

MR is not specific to treatment; even screening and other systemic interventions have been reversed, for example prostate cancer screening with the prostate-specific antigen (PSA) test (26) and glove and gown as physical barriers (27). As a phenomenon, MR has primarily been described in the field of medicine. However, a reversal may happen in any evidence-based field (12, 28). In this regard, Sutton et al. broadened the definition to the concept of evidence reversal (ER). In their words, ER occurs when an existing claim is tested, and the original evidence is contradicted by new and stronger evidence. The advantage of this definition is that it overarches all types of claims. In other words, claims supported by varying level of evidence, low to high quality, are susceptible to being contradicted by emerging evidence, with varying levels of susceptibility. This definition is more extensive as it states a claim as the starting point instead of an established medical practice. Furthermore, Sutton's definition includes a specific situation that does not seem to be fully incorporated in Prasad's definition of MR. Without a doubt, the definition of ER includes the more typical case where an existing claim supports a particular medical practice and new evidence contradicts this claim, showing for example that the practice is no better than doing nothing (15). However, an ER can also occur when the original claim is against a medical practice and new evidence emerges to contradict that claim, reporting efficacy of that medical practice. In fact, a study of reproducibility of clinical research in critical care

found that 4 of 35 (11%) reproduction studies with inconsistent effect can be described as follows: The original study reported a lack of efficacy while the reproduction attempt reported efficacy (29). In other words, an ER occurs when there is a change in the direction of effect supporting a claim due to new and better-quality evidence, regardless of whether the initial claim was supporting or opposing the practice.

As my project will investigate claims in the field of medicine, the definition of MR I will use is as follows: an MR is an ER in the field of medicine. In other words, an MR occurs when there is a change in the direction of effect supporting a claim due to new and better-quality evidence, regardless of whether the initial claim was supporting or opposing a medical practice.

Many concepts have been associated with MR (30) and it is important to differentiate between them. First and foremost, is the concept of replicability. This refers to the "ability of a researcher to [replicate] the results of a prior study if the same procedures are followed but new data are collected" (31). In science, replicability is one of the pillars of rigorous and reliable empirical research. Medical specialties such as critical care and thoracic surgery are tackling this issue by publishing frameworks and restating the importance of explicit protocols to increase replicability (32-34). This particular concept has become increasingly popular especially in the fields of biomedical science and psychology as the non-replicability of studies seems to be frequent, to the point of a 'replication crisis' (35-37). One notable example in the field of psychology is the effect of power-posing, holding specific postures that makes your body more physically expansive, on our subjective feeling of power and physiological factors (38). A larger study was later performed and could not replicate prior findings (39). A consensus on the effects of power-posing has yet to be reached (40, 41).

In clinical research, replicability can be seen as a practice being re-evaluated with a new set of participants and then evaluated for results and inferential reproducibility (29). Both results and inferential reproducibility are associated but do not systematically lead to an MR.

Results reproducibility is achieved when the results from the replication study – same method, different patients – corroborate the results of the original study (42). As for inferential reproducibility, it is defined as drawing the same conclusions from either a study replication or a reanalysis of data from the original study (42). On that note, research examining RCT data found that 13 (35%) of the published re-analyses could alter the conclusion of the original trial (43).

Although 35% may be an overestimate, re-analyses that contradict an original study bring uncertainty as to which conclusion should influence practice. Indeed, the original study will most probably be published in a more influential journal while the study involving a reanalysis, considered less valuable by some editors, may represent a better appraisal of the data.

Regardless of the definition of replicability, replication studies face two critical barriers. First, reproducing a study can be quite challenging. In the biomedical sciences, 10-25% of findings in the field are reproducible (36). In addition, replication studies are undervalued and therefore suffer from publication bias. This in turn causes the proportion of published replication studies to remain low, although growing (35). This is a problem for science as more replication studies are needed (29).

MR is not only closely associated with the concept of replicability; it could even be said that MR occurs in a subset of non-replicable studies. That being said, an important distinction to make is that non-replicability of a study does not automatically lead to an MR, and an MR is not necessarily the result of a reproduction attempt whose findings are inconsistent with the original study. In fact, non-replicability can lead to two situations: 1) A diminished (or augmented effect) but in the same direction of effect as the original claim, where a positive claim stays positive and a negative claim stays negative; 2) Medical reversal – new evidence contradicts the original research because the intervention in shown to be ineffective, or the direction of effect of the intervention has shifted between the new and the prior study. The proportions of non-replicable studies in situation 1 compared to 2 is not yet known. However, one study found that 24 (69%) of 35 inconsistent replication attempts lead to a change in the direction of effect. This further links the concept of replicability to the phenomenon of MR.

Although research regarding MR is sparse, the concerns surrounding MR are increasing and research in the field is gaining momentum. Lead by Prasad, several studies have now been conducted to investigate this phenomenon. In their first publication, Prasad et al studied the frequency of MR in all primary research of established practices published in the NEJM in 2009 (19). Of all 124 original studies investigating a clinical practice published that year, 35 (28%) were about a practice already adopted. The authors found 16 (13%) of these studies constituted a reversal. In other words, 16 of 35 (46%) original studies on medical practices already in adoption constituted a reversal. After this preliminary work, these authors extended the range of years of publication from one year to one decade (2001-2010) in the NEJM (15). They found 1344

original articles concerning a medical practice. Of these, 363 (27%) investigated an established practice, of which 146 (40%) reversed that practice.

In their latest publication, the group broadened their search to all RCTs regarding a medical practice published in three leading journals in a 15-year range. In this more recent work, they identified 396 (13%) overturned medical practices (44).

From these studies, a few lessons can be learned. First, MR remains a current and important issue in 2019. Indeed, across these three studies, 11-13% of original articles concerning any medical practice and 24-46% of original studies on already adopted medical practices were subject to a reversal. These numbers are consistent with a seminal publication reporting that 16% of original clinical research studies cited more than 1000 times were contradicted in subsequent studies (45). Second, a significant aspect about the method is this; the authors start with the newest study and determine whether the findings of that study reversed a medical practice. In other words, they consider the new RCT as the 'best' evidence and verify if what is done in practice is the same as the recommendation at the end of the new trial. In doing so, the authors identify medical reversals without looking at what evidence supported the established practice in the first place. This method helps to identify low-value medical practices (46-48) but does not differentiate between the varying quality of evidence supporting medical practices (49). On that note, evaluation of the evidence supporting medical interventions seems to show that only 11-13% of interventions are supported by good evidence (50, 51). Of course, not all evidence is considered to be equal. For example, an expert opinion is generally considered as lower quality evidence than conclusions drawn from observational studies and RCTs (52-54). For that reason, the set of MRs identified by Prasad and colleagues could have equally been from practices of unknown effectiveness as from practices likely to be beneficial, ineffective or harmful.

On the other hand, in my research, MR will be identified within RCTs only. New trials will be compared to previous RCTs and then I will explore whether any of the characteristics of these trials are indicative of future reversal. As mentioned before, MR has been associated with surrogate end-points, comparisons with the wrong controls, overconfidence in the role of pathophysiology, and extrapolating study results to an age group not studied (11). In doing so, my long-term objective is to provide physicians with a prediction tool that helps decide whether

the recommendation of a trial should be applied in practice or not, based on its potential to eventually be reversed. Before building such a tool, this research is necessary.

An important aspect to note is this: we have no perfect way to document the frequency of MR (55), therefore the proposed study is an approximation. In fact, the method is unique but greatly inspired by the step in a systematic review where two independent reviewers identify the set of relevant publications (56). Although we have no indication of the proportion of MRs in RCTs specifically relevant to primary healthcare, there should be fewer MR in RCTs than in what has been found in the literature. As previously stated, there is a higher chance of MR happening when clinical practices are not well supported by evidence; logically, MRs should have less chance of happening in well supported medical practices.

2.2 Reasons for selecting RCTs as our unit of analysis

In the present study, evidence from RCTs will be scrutinized to reveal MRs. Contrary to the seminal work on MR by Prasad et al. (11, 15, 19, 44), the starting point of this research is not an established medical practice, but rather an RCT. In doing so, there is no need to examine what is established in practice, which could be a very challenging endeavor. A focus on MRs at the level of the evidence helps to avoid the issue of variation in practice and the time it takes for new evidence to be integrated into practice (57). My assumption is that in general, recommendations from relevant RCTs are eventually applied in practice. Thus, an MR at the level of the evidence should enable a change in practice, a de-implementation or de-adoption of a practice.

There are many reasons for selecting RCTs over observational studies or other types of evidence. First, RCTs are considered the best study design to measure the effectiveness of a new intervention or treatment. They are the gold standard (58). Indeed, with an appropriate randomization minimizing or avoiding bias, the treatment effect is only attributable to the intervention itself (59). In addition, the choice of trials as our unit of analysis helps to optimize the impact of this research. Since RCTs generate the most reliable evidence in primary research, they constitute the evidence used in systematic reviews and meta-analyses. Consequently, evidence from RCTs are ideally what practices and guidelines should be based on, under the assumptions of the EBM model (60). In other words, experimental evidence is arguably the most influential for clinical decision-making. However, not all RCTs are of the same quality (54). While RCTs need to be well planned and carried out, they can be affected by many biases (4).

Furthermore, high-quality evidence does not seem as robust as we would like to believe. Across time, even high-quality RCTs once supported by robust evidence can be reversed, further proving the fluidity of evidence (61). For example, for many years, aspirin was used in the primary prevention of cardiovascular disease (62, 63). Yet, the effectiveness of aspirin is now challenged by the ARRIVE trial, showing that the purported benefits no longer seem to outweigh the risks (64). This shift in the risk-benefit balance was associated not with aspirin itself, but with external factors, namely changes in the population risk for cardiovascular disease, and increases in adjunctive medical therapy that itself reduced the risk of disease. Another reason for choosing RCTs is that due to their position in the hierarchy of evidence, each trial concludes with a recommendation on the intervention tested. As evidence from new trials emerge, provided the trial is well designed and carried out properly, these recommendations often become the new standard of care.

2.3 Appraising the quality of a randomized controlled trial

While no research has yet reported on the association between trials and the phenomenon of MR, many publications have investigated the impact of poorly conducted trials on the variability of treatment effect.

With the objective of investigating factors associated with an outcome of interest, I looked at different tools to appraise the quality of trials. Multiple scales and checklists have been devised to evaluate the likelihood of bias in a particular study or to multiple trials for comparison.

The first tool I examined was the *Jadad* scale. This consists of three closed questions concerning methodological aspects of the trial, namely [1] randomization, [2] double blinding, [3] withdrawals and dropouts (65). The scale rewards 1 point to positive answers with no partial points for a maximum of 3 points, equivalent to the best grade. This is a very simple and straightforward scale. That being said, the Jadad scale is very unforgiving for clinical settings where randomization and/or blinding is unethical or impossible (66).

For a more multidimensional tool, the Cochrane risk of bias (RoB) tool was also considered (60). This tool helps to evaluate the risk of bias of a trial, rather than the reported characteristics. The risk of each domain of bias – selection bias, performance bias, detection bias, attrition bias, reporting bias and other bias – are described as low, high or uncertain by the

evaluator. In an example given in reference 56, the authors represented each domain of bias previously mentioned as random sequence generation, allocation concealment, blinding of participants and personnel, blinding of outcome assessment, incomplete outcome data, selective reporting and other bias. These are very informative for data that can be extracted for my research. Still, the RoB tool may not be the most comprehensive approach to assess the potential for bias in randomized trials.

In the spirit of enriching our search, the CONSORT checklist was also examined. CONSORT is a checklist of recommendations for reporting randomized trials (67). By using this tool, adequate reporting and study design should be promoted, thus improving the ability of readers to judge the reliability and validity of a trial. Usage of this checklist has been associated with better quality trials (68). Even tools for the lay public, such as patients and their families, mention blinding of participants and of outcome assessment. As an example, the Informed Health Choices (IHC) framework lists 36 concepts the public should know about, to assess the trustworthiness of treatment claims and for decision-making (69).

The information from the tools mentioned was compiled with empirical evidence of trial influence to find characteristics of trials that could potentially be associated with MR. The quality of a trial is linked to the integrity of its randomization. Indeed, if the randomization of a trial is compromised, the estimated treatment effect may be misleading due to biased allocation (70). The two key elements that ensure the quality of randomization are the generation of random allocation sequence and the use of a proper allocation concealment method. Both of those are necessary to reduce selection bias (70, 71). In theory, by using an unpredictable allocation sequence, the randomization will generate similar groups that can be compared. In addition, by balancing known and unknown confounders, the process of randomization enables researchers to claim that an observed effect is directly associated with the intervention (72).

To be effective, the allocation sequence must remain a secret, hence the need for concealment. Allocation concealment are the steps taken to prevent clinicians and participants of being aware of which trial arm the patients will be allocated to (73). When allocation is not concealed, studies have shown that poor randomization leads to exaggerated intervention effects (74-77). That being said, it is theoretically possible for an improper allocation concealment to result in an underestimation of the effect of the intervention (71). Inadequate allocation concealment can also lead to deciphering of the allocation sequence, further influencing the

quality of randomization (71). A study found that close to a fifth of recently published trials in major medical journals used inadequate concealment and about a quarter failed to give a clear description of their allocation concealment method (77). The responsibility to ensure trials properly conceal allocation falls to the people involved in the trial; this includes researchers and participating clinicians. In addition, those involved in the publication of the study, which include journal reviewers and editors can also help to ensure biased trials are not published (78).

Blinding, also referred to as masking, is also important as it helps to reduce ascertainment bias (79). Blinding helps to mitigate unwanted bias and thus should be used when possible (34). On that note, an important difference between blinding and allocation concealment is that while some interventions cannot be properly masked, concealing the allocation can always be implemented into the design of an RCT (71). Trials not reporting double-blinding are associated with larger treatment effects than double-blinded trials (75). For example, the lack of or unclear double-blinding has been associated with a mean of 13% exaggeration of intervention effect (76). In another vein, sample size has also been linked with the probability of future contradiction, in that studies with small sample size can be refuted by larger studies (80, 81).

More importantly, while biases and other factors can lead to an under or overestimation of the treatment effect, it does not necessarily imply they will lead to an MR.

In some cases, an RCT may be ethically sound but infeasible, due to difficulties with randomization, blinding, and/or recruitment as with rare conditions. Another important limiting factor is the high cost of this design, both in terms of time and money, thus requiring careful consideration. While the study population of my study will be a subset of all randomized trials relevant to primary healthcare, selecting this sample of studies is neither simple nor straightforward.

2.4 The use of evidence by physicians

Under the EBM model, the evidence is a fundamental and constantly changing element. As mentioned before, integrating the best evidence in practice is a difficult task(82). Physicians in primary healthcare conduct their work under immense time pressure. Indeed, the time constraints they experience make the reading of new evidence a challenge of itself. In addition, the number of studies published is too much for anyone to handle (83). Moreover, physicians tend to lack the skills to critically appraise original research (10, 84). This makes it particularly difficult to

reconcile conflicting evidence. Journal reviewers and guideline producers also have their share of responsibility when it comes to publishing properly crafted recommendations.

One solution to circumvent some of these barriers is summaries of pre-appraised evidence (53, 85). For example, physicians will preferentially read synopses – concise descriptions of a single RCT or a systematic review – instead of the entire research study (86). Online information resources, such as DynaMed and Essential Evidence Plus (EE+), provide summaries of the evidence (87). One version of pre-appraised evidence is Patient-Oriented Evidence that Matters (POEMs). POEMs are succinct descriptions of recently published research, either an individual study or a systematic review. They are written for the purpose of educating clinicians (86). In the USA, POEM writers scan 102 journals to appraise more than 3000 studies monthly. They then select the ones that will be turned into POEMs (88). Similar to an abstract, POEMs provide a very brief overview of study design and results. POEMs differ from abstracts in terms of content. An example of a POEM can be seen in Appendix I. Arguably, the most important element is a "bottom line" statement summarizing the findings of the study. This is where the value of POEMs lies as it "is designed to help clinicians understand how to apply the results" (89). In addition, the information in a POEM about study design and results is provided to illustrate that an assessment of study validity was done. On another note, POEMs include the level of evidence (LOE) from the Oxford Centre for Evidence-based Medicine (90), a description of any financial support and a link to the PubMed entry for that study (88).

A particular advantage of POEMs is their relevance (91). The selection of studies to become POEMs has a built-in step to maximize the usefulness of this information to patients. In this case, usefulness is defined by three factors as per the following equation: $Usefulness = \frac{Relevance \times Validity}{Work}$ (92); where relevance includes the type of outcome (patient-centered or surrogate), the feasibility of the intervention for practice, the frequency of that clinical problem, and the anticipated impact on practice. Here, validity refers to the methodological rigor of the study. The work factor represents the time, money or effort required to obtain the information. In other words, information of high relevance and validity for minimal work is highly useful.

POEMs are sent out, one synopsis a day, by E-mail to clinicians on weekdays; hence the name Daily POEM. Because of their brevity and spacing over time, clinicians can more easily keep themselves updated. This may increase the chance of integrating POEM evidence in clinical practice. In Canada, POEMs can be used for continuing medical education (CME). Since 2006,

physician members of the Canadian Medical Association (CMA) can rate the Daily POEM for CME credit. Each POEM is archived online in a database that comprises more than 5700 POEMs. EE+ is the specific platform where all POEMs are made available for retrieval (93).

2.5 Reasons for choosing Daily POEMs

In the context of my study, the unit of analysis will be RCTs written in the form of POEMs, henceforth referred to as POEM-RCTs. The main reason for choosing Daily POEM-RCTs was to maximize the impact of this research. Since POEM-RCTs are commonly read by physicians, and RCTs are high on the pyramid of evidence, findings of a study investigating the rate of their reversal will be relevant. As a brief reminder, the focus of the study will not be at the level of practice, but rather at the evidence level. More specifically, the evidence available is appraised, regardless of whether or not a given therapy is implemented. That being said, about 25,000 CMA members receive the daily POEMs and the top 20 POEMs of 2018 have been referred to as "practice changers" in an upcoming conference (94). Therefore, the assumption is that at least some CMA members integrate evidence from POEMs into their practice. In addition, we chose specific evidence resources to mirror what clinicians following the EBM process would do in actual practice. In other words, many physicians read daily POEMs as a way to keep updated, and some may later use an online resource such as EE+ to retrieve that information, when needed at the point of care (95).

While confirming the Daily POEMs is truly used by physicians in their practice may be difficult, an assumption can be made that physicians using this resource integrate something from it, in their practice. To my knowledge, no research about MR for practices supported by evidence from RCTs in primary care and no research on MR of POEMs has been reported. However, at least one example of MR related to POEM-RCTs is known. In 2002, a trial concluded that duct tape seemed effective for warts (96). In 2007, two new RCTs contradicted this 2002 POEM-RCT by showing the ineffectiveness of duct tape for warts in children (97, 98). Subsequently, many clinicians stopped recommending duct tape as a treatment for the common wart.

2.6 Harms

The importance of the current study lies in the following statement: Any medical practice being reversed is associated with multiple avoidable harms (11). First, any intervention may cause harm to patients. Indeed, all drugs, surgeries, or other therapies have varying risks and side effects. Although side effects do not happen systematically, giving unnecessary treatment to a patient can directly affect health. In addition, all medical practices have a cost. This cost includes the price of the treatment but also that of other resources, such as human, financial, and organizational. In other words, MRs are associated with unnecessary costs. As for reversing a practice once it is shown to be ineffective, a commonly used practice can be difficult to de-adopt. In fact, robust and consistent evidence discrediting a practice may not be enough, as there is always some resistance to change. For example, findings of the COURAGE trial (99), which showed percutaneous coronary intervention (PCI) to be ineffective, caused an immediate reduction in PCI. Subsequently, editorials supporting PCI and case studies with positive outcomes were published in support the efficacy of PCI. Over time, the number of PCIs has gone back up, similar to the numbers prior to the COURAGE trial (100). Therefore, having good evidence against a practice might not be enough to stop it, even if a framework to guide the process of de-implementation exists (100). Furthermore, patients may lose trust in the health care system. In a situation where a patient has been recommended a treatment and is then later told it does not work, it is understandable that the doctor-patient relationship may be affected to the point where the patient may lose trust.

Before moving forward, note that my assumption is that physicians want to offer the best care to their patients. Managing the constraints, such as time pressure, and the uncertainty related to interventions is very difficult. In the case of a reversed medical practice, however, physicians share the responsibility for de-adoption.

There are no empirical estimates of the impact of these harms as they may differ from one practice to another and from one patient to another. Having said that, these consequences could be avoided with a more rigorous process for integrating trial results into practice. For this reason, studying the association between the characteristics of RCTs and MR may help to develop better guidance on when to adopt new practices.

2.7 Research Questions

Given the findings of my literature review, I want to know the following: First, how often are POEM-RCTs reversed? In addition, to better understand the phenomenon of MR and because I am unsure how many MRs we will find, my secondary question is as follows: what factors, here characteristics of RCTs, are associated with MR in primary healthcare?

III. Methodology

3.1 Overview

My literature review revealed no standard method to document medical reversals. For this reason, we (Drs Grad, Shaughnessy, Ebell, Slawson, and myself) devised a method that is greatly inspired by the selection of relevant studies in a systematic review. In a nutshell, the evidence from an awareness service, here POEM-RCTs published in years past, and subsequently the evidence from DynaMed will be compared (82).

The method is composed of 2 steps. Step 1 consists of identifying POEM-RCTs that are medical reversals and then describe the shift in the direction of intervention effect. Step 2 consists of extracting trial characteristics that are potentially linked to the outcome of interest, here MR.

The data analysis is comprised of descriptive and inferential statistics. The latter involves conventional regression modeling (logistic regression) as well as modern approaches including LASSO regression, classification trees and random forest to identify and rank statistically important predictor variables for MR.

3.2. Selection of POEM-RCTs

Mark Ebell, professor at the University of Georgia, sent my supervisor a database (Microsoft Excel (version 16.24) spreadsheet) of all POEMs ever disseminated, in September 2017. The first inclusion criterion concerned selecting POEMs with the study design of interest, here RCTs. More concretely, all POEM-RCTs in the database that had the following labels as study design were included: “Randomized controlled trial”, “Randomized controlled trial (double-blinded)”, “Randomized controlled trial (nonblinded)”, “Randomized controlled trial (non-blinded)”, “Randomized controlled trial (single-blinded)”, “Randomized Controlled Trials”. In other words, POEMs on observational, ecological and qualitative studies, reviews, practice guidelines, etc., were filtered out. Crossover trials were excluded as this design was considered different from RCTs especially when it comes to their sample size, as the sample size of crossover trials tend to be smaller (101, 102).

In addition, when deciding which years to focus on, I wanted to allow sufficient time for new evidence to emerge after the publication of the original research on which the POEM is based (61). For this reason, I chose a minimum of 10 years between the time the POEM was published and the start date of this study. Ergo, the latest year the POEMs were taken from was 2007. The aim was to finish with a subset of about 1000 POEMs. The POEM-RCT inclusion process is illustrated in Figure 1. Ultimately, of 5738 POEM records available to us, the POEM-RCT dataset we selected was a more manageable subset of 960 POEM-RCTs from 2002-2007.

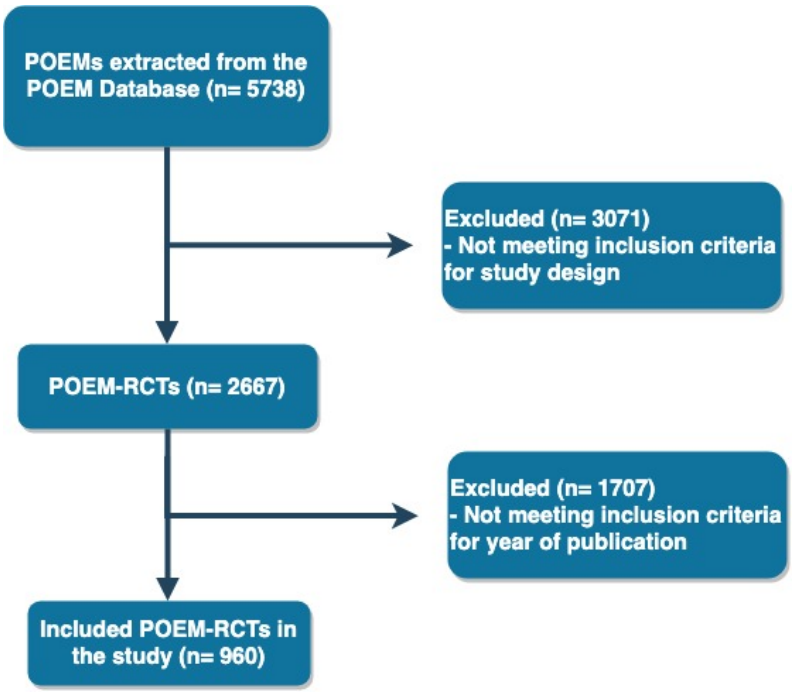

Figure 1. Flow Chart – Selection of POEM-RCTs

3.3 Data collection

As briefly described above, data was collected in a two-step process from September 2017 to March 2019. During this process, I supported raters and answered their questions, when needed.

3.3.1 Step 1

The objective of this step was to identify the outcome of interest, whether the bottom line of an RCT was reversed or not, in the subset of POEM-RCTs. In addition, whenever a reversal was

identified, details about the shift of the direction of effect was recorded. In order to do so, raters were recruited.

3.3.1.1 Recruitment of raters

With the help of my supervisor, I individually contacted physicians from the United States who were familiar with the Daily POEMs. Recruitment started in September 2017 and finished in December 2018. Four raters were initially recruited from McGill, Tufts University, University of North Carolina at Chapel Hill and the University of Georgia. After the initial four, snowball sampling facilitated the recruitment of 5 other raters, for a total of 9 raters. These raters had between 3 and 35 years of experience and had one of the following degrees: MD, MD CM, MBBS, DO, and PharmD.

Once raters agreed to participate, two documents were provided to them. Document one supported the coding process in Step 1. Document two provided further information for them to complete their task. Both documents are available in the Appendix II and III, respectively. These documents were prepared for two main reasons. First, a properly designed codebook with appropriate definitions was necessary to standardize the coding process. Second, both documents offered support during the coding process and clarified the role of the rater. Having such clarity helped with the recruitment of raters.

3.3.1.2 Document 1 – Operational Definition and Coding Guide

The purpose of this document was to standardize the coding process in Step 1 by providing each rater with operational definitions of MRs and the types of MRs, examples, and explanations. Indeed, raters needed to better understand the concept of MR.

The document included an operational definition of MR and of the types of MR. Essentially, this document provided explanations of the two elements collected in Step 1, and a codebook. The codebook was comprised of a description of the coding template (see *Document 2*), the detailed step-by-step description of the coding process, and the coding guide, providing an explanation of each code.

This was the first document provided to each rater. Each rater was asked to read it and ask questions, as needed. The nine raters were also asked to follow the coding process, to standardize coding in Step 1.

Document 1 was developed through an iterative process. The first version was shown to the 4 original raters. Through feedback and revisions during four online meetings and a trial run to rate a small number of POEMs, the information in the document was consolidated and clarified. More details on the trial run will be given in the next subsection.

3.3.1.3 Document 2 – Coding Template

The coding template was a document providing information for raters to complete Step 1. It is also the document where the outcome of interest, MR, and the type of reversal are reported. Originally, the template was in a *Word* document, but was changed based on feedback from the raters following the trial run. In fact, since the final study database was in *Excel*, using the same template for data collection in Step 1 optimized and facilitated the integration of POEM-RCT ratings. In other words, this process reduced the risk of human error in entering the data.

Each coding template contained 50 POEMs from the study subset of 960. This number of POEMs was arbitrary and selected so as not to overwhelm the raters. Furthermore, the document included the full POEM (title, clinical question, bottom line, reference, study design, setting, synopsis, POEM identification number), a DynaMed statement (E_2), and a column for comments (see Appendix III). The bottom line of the POEM is a recommendation that reflects the most important take away message for physicians. This bottom line corresponded to the initial evidence (E_1) published 2007 or earlier. The DynaMed statement (E_2) was the updated evidence I extracted from DynaMed in 2019. This E2 was used by raters as a point of comparison to E_1, to identify MRs. More information about E_2 will be given after discussing a trial run done with the initial 4 raters, below.

To test out, adjust and improve the data collection process for Step 1, a trial run of 10 POEMs with the 4 initial raters was completed. From this trial run, modifications to the presentation of E_2 were made. Indeed, E_2 was originally extracted from DynaMed verbatim. However, this method could be very lengthy. More concretely, in a case where the same evidence was found for E_1 and E_2 or when E_2 was consistent with E_1, E_2 would have been lengthy to read. This would have greatly reduced the efficiency of data collection. For this reason, by mutual agreement between with raters, we decided the most efficient way to present E_2 was by using different statements depending on the situation. I called these the DynaMed Statement.

The different statements were as follows:

- When the intervention was not mentioned in DynaMed, I would state: *the intervention was not mentioned in DynaMed*;
- When the evidence found in DynaMed was the same as the bottom line of the POEM, either because no new evidence has been found or because some new evidence was consistent with the POEM, I would state: *the intervention was mentioned in DynaMed and no contradictory evidence was found*;
- When the evidence found contradicted the bottom line of the POEM, I would state: *the intervention was mentioned in DynaMed and the following contradictory evidence was found* « [Insert text verbatim from DynaMed] »;
- When I was unable to determine if the information found was similar or opposed the bottom line of the POEM, I would state: *The intervention was mentioned in DynaMed and the following evidence was found; however, I am unable to determine if this is contradictory evidence* « [Insert text verbatim from DynaMed] ».

For each POEM-RCT, I restricted my search in DynaMed to a 15-minute time-limit for E_2. For practical reasons, this time limit was imposed as in some cases, such as for alternative therapies, the interventions were not mentioned in DynaMed. Whenever this happened, I would state: *I was unable to extract the evidence specific to the POEM after a 15-minute search. I recommend raters search the literature*. Moreover, raters were invited to seek further evidence from another resource, such as *UpToDate*; however this was not a requirement. The rationale for this was the following. As in clinical practice, physicians use their knowledge and the evidence in a process of decision-making. Given the time constraints of clinical practice, they decide on the number of knowledge resources they consult.

3.3.1.4 Rating Process

After reading the operational definition and the coding guide, raters were asked to complete the coding template. In order to do so, a rater received a coding template with a set of 50 POEM-RCTs. S/he would read the information from the POEM, E_1, E_2, and any personal note I included for the raters. For example, for a particular POEM (103) my personal notes were: “New evidence, not sure if consistent with the bottom line of the POEM. The bottom line seems to

have two results, both being isoflavone but from either red clover or soy. For this reason, I extracted two different pieces of evidence". Then, the rater would compare the initial (E_1) and updated (E_2) evidence of efficacy and answer the following question: Is the bottom line of this POEM (E_1) reversed in 2019?

There were four possible answers encompassing all possibilities.

(0) **No**. When E_1 is not reversed in 2019. In this case, the updated evidence E_2 is consistent with E_1.

(1) **Yes**. When E_1 is reversed in 2019. Here, the updated evidence E_2 is inconsistent with E_1, suggesting a shift in the direction of effect.

(2) **Uncertain or cannot be resolved**. When a POEM contains many outcomes with different or uncertain shifts in the direction of effect.

(3) **No and cannot be resolved**. When E_1 is not reversed in 2019 and the situation cannot be resolved. This fourth option was added after a rater stumbled upon a very specific situation where the effect of omalizumab is discussed. In this specific case, as no subsequent studies have been published, according to DynaMed, E_1 would not be considered as reversed in 2019. Having said that, since the drug was taken off the market, E_1 can never be reversed, and it is a situation that cannot be resolved.

In addition, I wished to collect some information to be able to describe the type of reversals that were identified. The type of reversal here is used as the type of shift in the direction of effect between E_1 and E_2. More concretely, in a case of an MR, was it a situation where E_1 supported a medical therapy while E_2 showed it was ineffective or harmful? Or vice versa. Therefore, when E_1 was found to be reversed in 2019, the raters were asked to add another piece of information.

With respect to MRs, there were seven coding possibilities:

(0) **No reversal.** When the E_1 is still valid in 2019. This was added so there was no missing data in the dataset.

(1) **Strong reversal.** When E_1 strongly supports a practice and E_2 is strongly opposed to the same practice.

(2) **Weak reversal.** When E_1 strongly supports a practice and E_2 is weakly opposed to the same practice.
(3) **Weakest reversal.** When E_1 weakly supports a practice and the E_2 weakly oppose the same practice.
(4) **Strong reversal.** When E_1 is strongly opposed to a practice and E_2 strongly supports the same practice.
(5) **Weak reversal.** When E_1 is strongly opposed to a practice and E_2 weakly supports the same practice.
(6) **Weakest reversal.** When E_1 is weakly opposed to a practice and E_2 weakly supports the same practice.

Two situations were omitted from the coding process, by accident. Here, the two situations were the following: 1) When E_1 weakly opposes a practice and E_2 weakly but strongly supports the same practice; 2) When E_1 weakly supports a practice and E_2 strongly opposes the same practice. When this happened, the raters wrote them down as "-1 to +2" and "+1 to -2" directly in the coding template.

The types of reversals were subsequently sorted into two groups (same or opposite direction) depending on the change in the direction of effect between E_1 and E_2.

As previously mentioned, this rating process was inspired by methods to classify articles for a systematic review of the literature. In this regard, each POEM-RCT was categorized independently by two raters. Then the raters sent their completed coding template to me. After receiving these codes from each pair of raters, I compared their ratings. Any disagreements were identified as discrepancies and returned to them. First, an attempt to resolve these discrepancies by discussion was made between rating pairs. When an agreement was reached, it was considered as a final decision. When a disagreement persisted after this discussion, a third reviewer was invited to resolve this discrepancy. The rating process, as well as the number of agreements and disagreements were tracked.

3.3.2 Step 2

The objective of the second step was to extract the characteristics of trials summarized by each POEM. I used qualitative content analysis for this data extraction step (104). When extracting

trial characteristics, I started with the POEMs information. If any data was missing, I went back to the original trial to complete this data extraction. In addition, various data points were extracted directly from the EE+ database – study design, concealment, LOE, year, setting, and supertype. The POEMs database is developed and sustained by the POEM authors; their data extraction process is manual.

3.3.2.1 Study Design

All POEM-RCTs are classified by the POEM authors as double-blinded, single-blinded or non-blinded RCTs. Thereby, this data point gave information on the blinding involved in each trial.

3.3.2.2 Allocation concealment

As mentioned before, this information was extracted from the EE+ database. In conversation with a POEM author, Dr. Allen Shaughnessy, I learned how they extracted the allocation concealment status of a study.

During the process of writing POEMs, the authors evaluate the methodological quality of each trial, including whether allocation concealment was reported. When in doubt, the POEM author contacts the authors of the trial to get more information on the strategy for concealing allocation. Allocation concealment status is then documented in the 'synopsis' section of the POEM or in a specific column for this data, introduced in 2004 in the EE+ database.
Three categories were possible for allocation concealment.

(0) **Unconcealed.** When allocation concealment was inadequate. This tends to happen on rare occasion as the authors would have to clearly admit to not masking the allocation and/or to manipulating the process;

(1) **Concealed.** When the allocation concealment was adequate. In this case, the authors clearly described how they assigned participants to each group in a scientifically rigorous manner;

(2) **Uncertain.** When the allocation concealment is unclear. In this occasion, either the allocation was concealed properly but reported poorly, or the allocation itself was inadequate leading to a poor reporting.

3.3.2.3 Level of Evidence (LOE)

The LOE is a grading system developed by the Center for Evidence-Based Medicine at Oxford University (90). This provides a description of the quality of the evidence.

3.3.2.4 Year

This data point corresponds to the year the POEM was published. Most POEMs are published in the same year as the article they summarize. However, the original study may have been published in a prior year, especially for POEMs published at the beginning of each calendar year.

3.3.2.5 Setting

There are 11 different settings in the EE+ database, grouped into the 5 following categories: 0- Inpatient; 1- Outpatient; 2- Emergency department; 3- Population-based; 4-Other.

3.3.2.6 Age Group

The age group has three categories describing the age of the population in the RCTs. The categories were as follow: 1) Adults; 2) Children; 3) Both.

3.4 Sample size

This study characteristic was extracted in three parts to be as descriptive as possible. These three data points were the total sample size of the trial, the sample size of the intervention group, and the number of trial arms.

3.4.1 Total sample size

The total trial sample size is defined as the number of persons enrolled in the trial that underwent randomization, thus after eligibility criteria are applied and before any loss to follow up.

Coming up with the operational definition for this item was an iterative process. The definition was in fact modified during the extraction process for two reasons. First, each trial did not necessarily report on their total sample size in the same way and some are not as clear as others. Second, some protocols require a particular step in between the recruitment of patients and their randomization. For example, some trials have a preliminary phase prior to

randomization (105) where some participants may be lost or excluded. For these reasons, the total sample size I extracted is the total number of people randomized, regardless of the number who consented at the beginning of the trial.

3.4.2 Sample size intervention group

In most cases, this is very straightforward as it is the number of participants randomized to the intervention before loss to follow-up. That being said, extracting the sample size of the intervention group was more complicated on occasion.

For example, a trial compared the effect of ximelagatran to warfarin for the prevention of venous thromboembolism after total knee replacement (106). Although there is no placebo in this trial, warfarin was the standard therapy and is thus not considered to be the intervention. What is more, two of the three arms in the trial used ximelagatran at different dosage. For this reason, I calculated the sample size of the intervention group here to be the mean of the two ximelagatran groups. In another example, researchers compared the effect of ginkgo biloba and acetazolamide for preventing acute mountain sickness among Himalayan trekkers (107). In this trial, the compared treatment courses are 1) ginkgo, 2) acetazolamide, 3) acetazolamide & ginkgo, and 4) placebo. Since both ginkgo and acetazolamide are treatments under investigation, the size of the intervention group in this case was the average of all three active drug groups. There were yet other examples of POEM-RCTs comparing different doses of the same drug (108, 109). In cases where a trial had multiple interventions or intervention groups without one group being clearly identified as treatment or placebo, the average was calculated as the intervention group size. In some POEM-RCTs, there was no specific control group. For example, one POEM-RCT compared ibuprofen to acetaminophen, both combined with codeine and caffeine, as treatment options for pain from episiotomy or tearing of perineal tissues (110). For this example, the intervention group was the average of the participants in both groups.

Both definitions – total sample size and sample size of the intervention group – were chosen to make extraction manageable and to extract the numbers that would be used in an intention to treat type of analysis, regardless of the type of analysis that was actually done in the RCT.

3.4.3 Number of trial arms

Here, the number of trial arms is defined as the number of groups to which people were randomized. It is usually equal to the number of different treatment courses that were being compared.

3.5 Information from the Research Question

From each POEM-RCT, I wanted to extract data about the research question as this provides context as to what the trial was about. For this reason, I extracted information about the essence of the clinical question and the nature of the medical therapy under investigation.

3.5.1 Essence of the research question

This data point was selected to provide some information about the research question. For this, an existing taxonomy to capture the essence of doctor questions about patient care was used (111). The taxonomy of John W. Ely and colleagues contains 4 hierarchical levels of specificity to categorize a total of 64 generic question types. In the context of my research, the broadest level of specificity was used to simplify the coding process. The five broad areas of the first level are as follow: diagnosis, treatment, management, epidemiology, and non-clinical.

3.5.2 Supertype of the research question

Similar to the essence of the research question described above, the supertype provides more information about the topic addressed by the research question. The supertype is a more detailed classification system used by POEM authors than the broadest level of Ely's taxonomy. This information was extracted directly from the EE+ database. A table with the details about each supertype can be found in Appendix IV. I grouped the data under the following supertype: treatment (Tx), screening (Sc), prognosis and follow-up tests (Px), diagnosis (Dx), and etiology (Et).

3.5.3 Drug or non-drug intervention

As we selected POEMs of RCTs, the interventions studied were often drugs but sometimes procedural, such as a type of surgery. Therefore, I extracted information from each POEM

concerning the nature of the intervention. This extraction was a binary code. The intervention could fall under a drug – defined as anything that can be swallowed, injected, inhaled, or applied on skin as a topical/transdermal product – or not a drug, for example the type of surgery.

3.6 Data Analysis

3.6.1 Descriptive Statistics

The POEM-RCTs were analyzed by group, depending on whether they were reversed or not in 2019. POEM-RCTs that were identified as uncertain or cannot be resolved and not reversed were excluded for the analysis. Proportions of trial characteristics were calculated and presented for comparison purposes. In addition, I counted the number of disagreements in Step 1 of the method and the type of reversals we identified.

3.6.2 Inferential Statistics

3.6.2.1 Data mining and variable transformation

Before moving forward, it is important to note the following. Several variables were transformed to facilitate the interpretation of the output of statistical models. The variables 'total sample size' and the sample size of the intervention group were combined under a new variable "sample size ratio", calculated as follows: $sample\ size\ ratio = \frac{Sample\ size\ of\ the\ intervention\ group}{Total\ sample\ size}$. The ratio ranges from 0 to 1 and the higher the ratio, the closer the size of the intervention group is to the total sample size. The purpose of generating this variable is to see whether the size of the intervention group relative to the total sample size is important as a predictor variable.

As for the total sample size, it was categorized in four groups: (1) (0 , 100) participants; (2) [100 , 250) participants; (3) [250 , 500) participants; and (4) [500 , 40 000) participants. The selection of these thresholds was informed by the quartiles of the distribution of total sample sizes. The number of trial arms were summarized in a categorical variable with three groups: (2) two-arm trial; (3) three-arm trial; and (4) any trials with more than three trial arms. LOE has been changed into two categories: (0) all LOE 1a, 1b and 1b-; (1) all LOE 2b, 2b- and 2c. In addition, the *supertype* variable was aggregated into three categories: 0) treatment [Tx]; (1) screening [Sc]; and (2) Other. Finally, the variable *design* was completely taken out because all the completed POEM-RCTs were randomized controlled trials (double-blinded).

3.6.2.2 Data modeling approaches

All statistical analyses were performed using the software R (112). All R commands are available in appendix V. The first data mining step was to remove cases where the POEM-RCTs were rated as uncertain (2) or not reversed and cannot be resolved (3). This way the outcome of interest (MR) was guaranteed to be binary, either reversed or not reversed.

In addition to reporting standard descriptive statistics, four statistical modeling approaches were applied: i) binary logistic regression analysis, ii) Least Absolute Shrinkage and Selection Operator (LASSO) regression, iii) classification trees, and iv) random forest. The objective of applying these models was to further explore variable associations in the data and to identify statistically relevant predictor variables for MR.

Based on the results of the binary logistic regression analysis, a nomogram was created to visualize the relative contribution of each included candidate variable for predicting the probability of reversal. Nomograms are often used for clinical decision-making e.g. in oncology and add value to improving the interpretation of statistical models for individual case prediction (113). Each variable of the multivariate logistic regression model is included in the nomogram and is assigned a specific number of points. The sum of all points corresponds to an estimate of the individual risk of medical reversal (114, 115).

Since the expected number of medical reversals identified from Step 1 was low relative to the number of trial characteristics (model variables), as will be reported below, the risk of overfitting the conventional logistic regression model will be taken into consideration. Overfitting is a statistical issue that arises when overly complex models are used to make inference from relatively sparse data, rendering model-based parameter estimates and predictions unreliable (116).

In order to address potential overfitting issues, LASSO regression methods are proposed in the statistical literature (117). I therefore complemented the logistic regression analysis with a respective LASSO regression analysis. This allowed us to enter all candidate predictor variables in the model despite the relatively low prevalence of the outcome. Subsequently, a classification tree was fitted to the data to explore possible variable interactions (combination of predictor variable levels) that were not incorporated in the logistic and LASSO regression models.

Finally, a random forest model was fitted that builds on classification trees, addressing overfitting and related inaccuracy issues through the incorporation of resampling algorithms.

Random forests have demonstrated superior performance compared to classic statistical inference approaches such as ordinary binary logistic regression and classification trees (118). In order to give equal weight to sensitivity and specificity when maximizing overall predictive accuracy through the random forest, the input dataset was weighted to emulate equal prevalence of medical reversals and non-reversals.

To assess and rank variable importance with regard to predicting medical reversals, the increase in classification error (overall predictive accuracy) due to removal of the candidate variable from the random forest was determined. To enable robust interpretation of these random forest results in the given limited sample size setting, two random variables were generated: *randvar*, a variable following standard normal distribution, and *randvar2*, a binary variable with prevalence 0.5. The computed variable importance for these two noise variables was used as benchmark when assessing the predictive importance of the candidate variables. The increase in prediction error due to removal of a candidate was primarily used to rank variable importance and not to interpret distance in terms of predicted capability.

The obtained variable importance rankings from the random forest were then compared to the most relevant variables depicted by the logistic regression model (illustrated through the nomogram) and the LASSO regression.

IV. Results

4.1 Descriptive Statistics

4.1.1 Selection of trials into the research study

The flow chart detailing the selection of trials into this research is presented in Figure 2. Of 960 POEM-RCTs from the POEMs database, data on 410 POEMs was extracted and analyzed. The evidence of one POEM-RCT (119) was separated into five independent data contributions as this publication summarized five trials. Indeed, this particular study compared patient satisfaction of rizatriptan with other triptans; however, the data were pooled across five different trials: three of these were parallel group trials while the other two were crossover trials. As such, this POEM-RCT was treated as five independent studies. As two of these five studies were crossover trials, they were excluded, in line with our selection criteria. In addition, three of the 410 POEMs were subsequently excluded; two were crossover trials (120, 121) and one was a placebo-controlled trial with no randomization (122). Another POEM-RCT (123) was excluded because the RCT on which it was based was "retracted due to acknowledgment of scientific misconduct resulting in concerns regarding data integrity and inappropriate assignment of authorship" (124). This brings the final total to 408 data entries of POEM-RCTs.

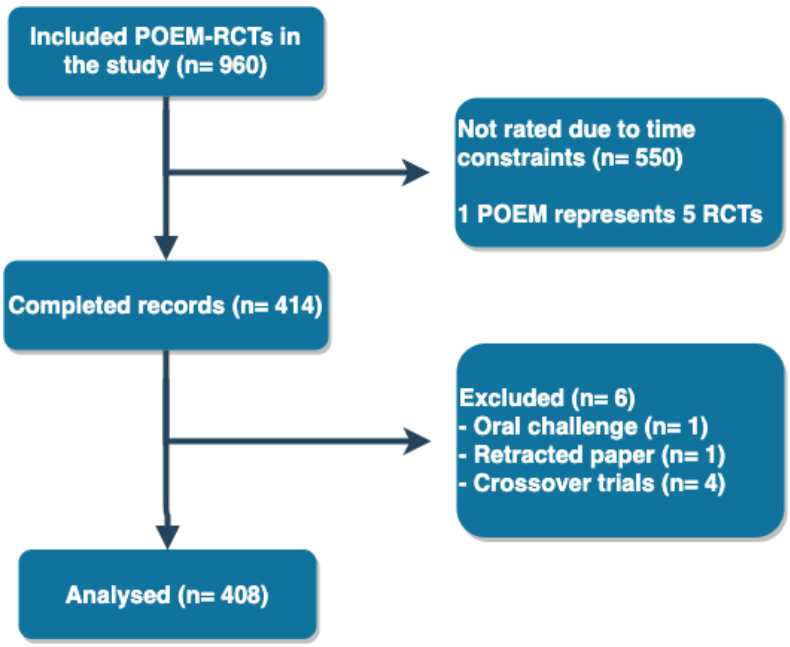

Figure 2. Flow Chart – POEM-RCTs Excluded or Analyzed

4.1.2 Primary outcome – Reversal

Of 408 POEM-RCTs, 34 (8% ; 95% CI [6 , 11.4])) were identified as reversed, 360 (88%) were identified as not reversed and 14 (~3%) were inconclusive. More precisely, 11 (~3%) were 'uncertain and/or cannot be resolved' and 3 (~1%) were identified as both 'not reversed' and 'cannot be resolved'. A description of all 34 reversed POEM-RCTs is provided in Appendix VI.

4.1.3 Types of reversal

The shift of evidence that resulted in the reversal of our 34 POEM-RCTs is equally split, from positive to negative and from negative to positive, as represented in the Figure 3. In other words, there are as many POEM-RCTs where E_1, the bottom line of the POEM, supports a medical practice while E_2, the updated evidence from 2019, does not as vice versa.

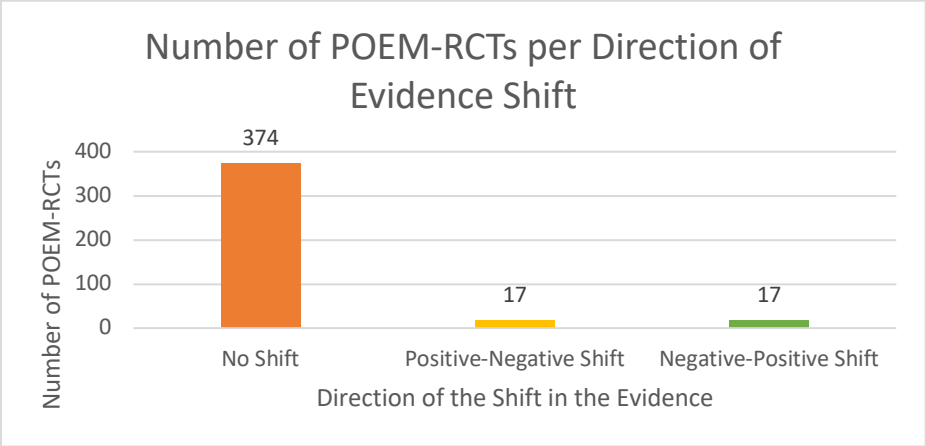

Figure 3. Direction of shift in the evidence

4.1.4 Disagreements

In total, raters disagreed on 67 of 408 (16%) POEM-RCTs in Step 1. Of those disagreements, 62 (93%) were resolved by discussion between the two independent raters. The remaining 5 (7%) disagreements were resolved by a 3rd reviewer.

Once resolved, 34 (51%) of the disagreed POEM-RCTs were identified as not reversed, 21 (31%) as reversed, and 12 (18%) as inconclusive, meaning either the potential outcome of reversal remained uncertain or not reversed + cannot be resolved. These results can be seen in Figure 4. The total number of inconclusive is of 14 as 2 POEM-RCTs were identified as uncertain by both raters.

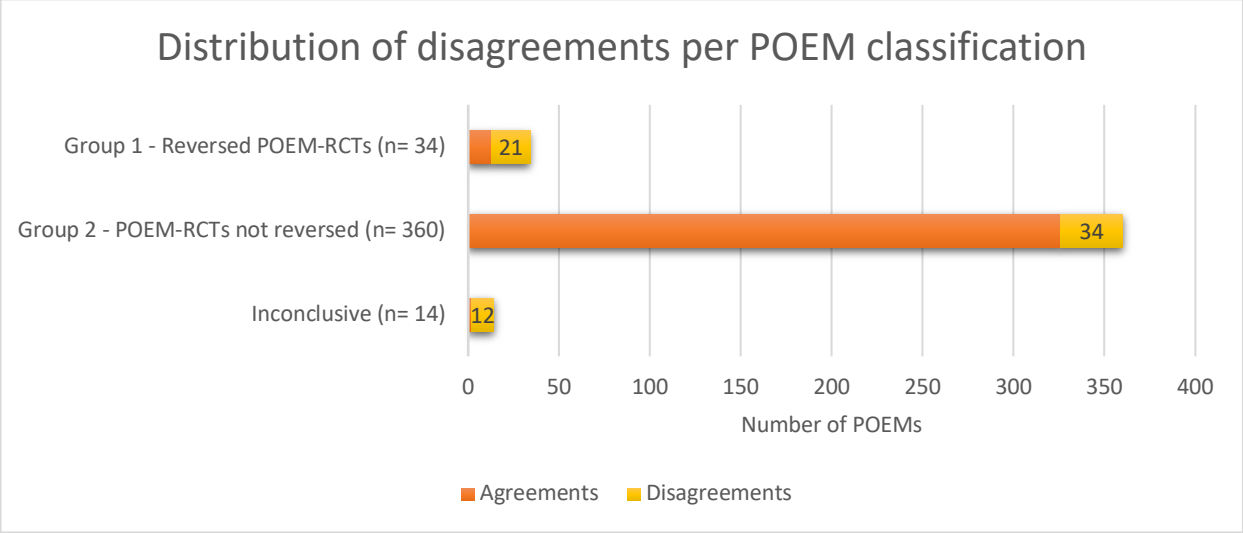

Figure 4. Distribution of disagreements per POEM-RCT classification

As seen in Figure 4, there is a proportionally large number of disagreements for ratings of POEM-RCTs in Group 1 and in the inconclusive group. This may be explained as follows: Our ability to recognize a medical reversal is not as a strong as we thought.

Based on the figure above, I believe that a measure of inter-rater reliability (such as Cohen's kappa) will not be helpful. Figure 4 shows that rater pairs often disagreed in classifying POEMs in Group 1 and in the inconclusive group.

4.2 Exploration of factors associated with reversal

4.2.1 Descriptive comparison of marginal characteristics of the subset of POEMs

All 408 POEM-RCTs are double-blinded RCTs from 2002-2005. The total sample size of these POEM-RCTs averaged 2,326 participants, ranging from 12 to 39,876 participants. The mean size of the intervention group was 1047 participants; with the intervention group ranging from 6 to 19,937 participants. The most common study design was the two-arm parallel design (76%), followed by three-arm parallel designs (14%), and four-arm parallel designs (9%). For these POEMs, the study setting was outpatient (73%), inpatient (16%), population-based (5%), emergency department (4%), and nursing home / rehab unit / other (2%). Most trials studied an adult population (86%). Two-thirds of the POEMs described allocation of participants to the respective study arms as being concealed. In the other third, allocation is uncertain.

About 93% of these POEMs summarized a trial investigating a drug. Under Ely's taxonomy (111), 401 (98%) POEM-RCTs focused on improving a treatment, and very few were about diagnosis. These results are very similar for the supertype classification system employed by POEM authors. Indeed, while 37 POEM-RCTs (9%) were categorized as pertaining to screening, 365 (89%) were about treatment. Just two POEM-RCTs were about prognosis and follow-up tests, three were about diagnosis, and one about etiology.

For comparative purposes, the study characteristics of two groups of POEM-RCTs are presented in Table 1. Group 1 consists of the POEM-RCTs identified as reversed. Group 2 are the POEM-RCTs identified as not reversed. The study characteristics of both groups seem to be similar. That said, on average, reversed POEM-RCTs have a lower sample size than group 2, while the proportions in allocation concealment also seem different, with group 1 having 79% concealed allocation and group 2, 64%. For all other variables listed in Table 1, no considerable differences between group 1 and 2 were apparent.

*Table 1. Characteristics of 394 POEM-RCTs **

	Group 1 - Reversed	Group 2 - Not Reversed
Number of POEM-RCTs	34	360
Number of POEM-RCTs per year		
2002	2 (6%)	112 (31%)
2003	11 (32%)	102 (28%)
2004	11 (32%)	77 (21%)
2005	10 (29%)	69 (19%)
Total Sample Size		
Mean	1870	2414
Standard Deviation	6811	5801
Median	265	337
Range	[39 ; 39,876]	[12 ; 39,876]
Size Intervention Group		
Mean	889	1082
Standard Deviation	3403	2639
Median	112	139
Range	[13 ; 19,934]	[6 ; 19,937]
Setting - n (%)		
Outpatient	25 (74%)	261 (73%)
Inpatient	3 (9%)	60 (17%)
Emergency department	3 (9%)	14 (4%)
Population-based	3 (9%)	18 (5%)
Other	0	7 (2%)
Age group - n (%)		
1 - Adults	26 (76%)	273 (76%)
2 - Children	5 (15%)	39 (11%)
3 - Both adults and children	3 (9%)	48 (13%)
Allocation Concealment - n (%)		
Concealed	24 (71%)	230 (64%)
Uncertain	10 (29%)	130 (36%)
LOE - n (%)		
1b	27 (79%)	272 (76%)
1b-	6 (18%)	44 (12%)
2b	1 (3%)	34 (9%)
Supertype - n (%)		
Treatment	30 (88%)	321 (89%)
Screening	3 (9%)	34 (9%)
Other	1 (3%)	5 (1%)

*Excluded POEMs were 11 uncertain and 3 cannot be resolved + not reversed

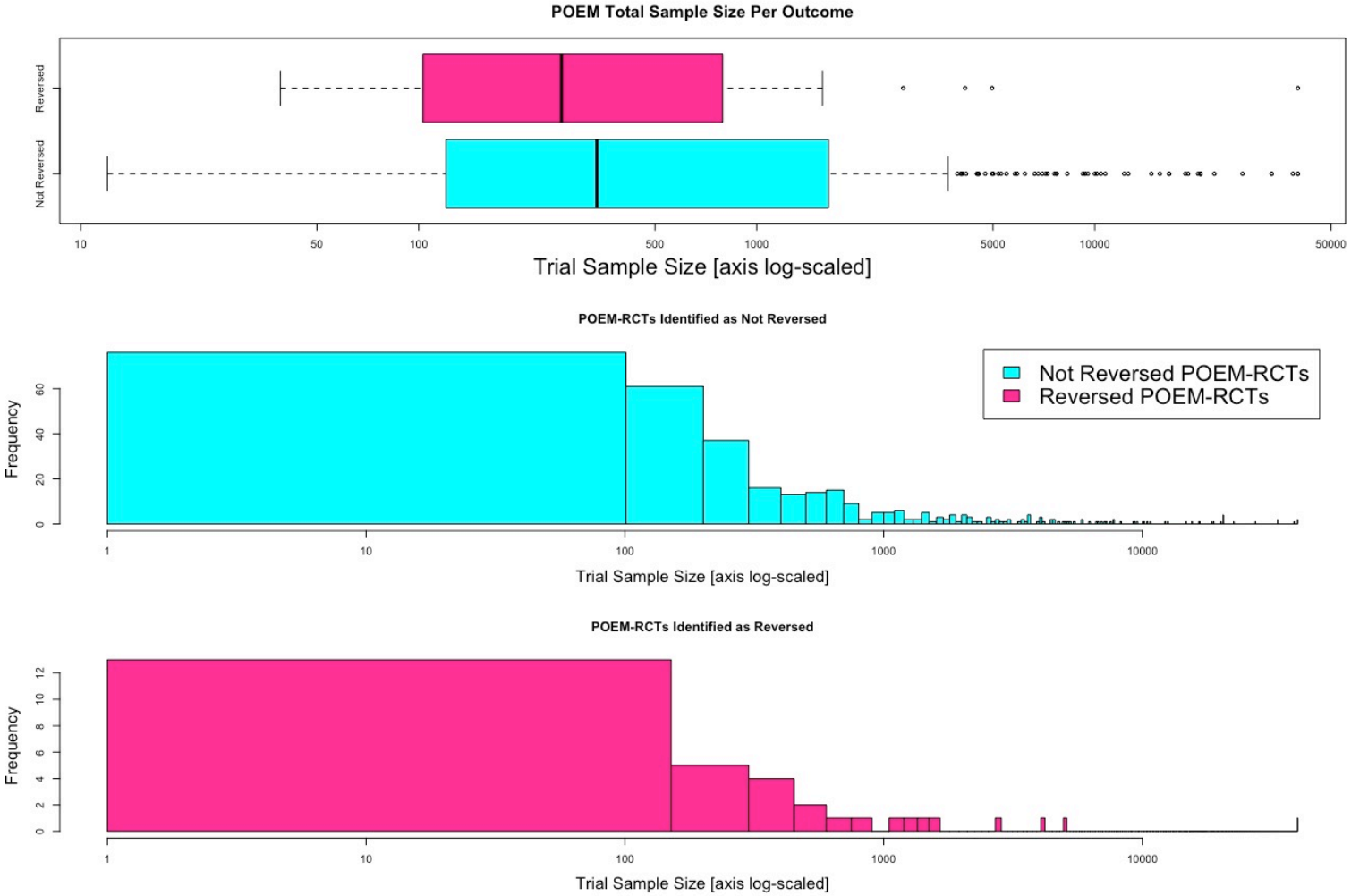

Figure 5. POEM-RCTs: Total Sample size per outcome

The data in Figure 5 indicated a difference in the distribution of sample sizes between trials that published study findings that were eventually reversed (Group 1) and trials that were not subject to reversal of study findings (Group 2). Indeed, besides the outliers in Group 1, most reversed POEM-RCTs have a total sample size under the 3rd distribution quartile of Group 2.

Further analyses of other factors associated with "reversal" of trial findings are presented in subsequent sections of this chapter.

4.2.2 Data modeling approaches to investigate associations

There were no major group differences with respect to the following variables: trial arms, supertype, age group, LOE, and setting. Nevertheless, I further assessed for conditional differences across the groups. For this purpose, different statistical modeling approaches were applied including binary logistic regression, LASSO regression, classification trees and random forests.

Neither the study design nor the specific research question topic were considered as relevant factors for these regression analyses. Indeed, all included POEM-RCTs were double-blinded RCTs and about 98% concerned the comparative evaluation of a new treatment or intervention.

4.2.2.1 Logistic regression analysis and nomogram

As outlined in the methods section, the first exploratory analysis for identifying study-level factors associated with the primary outcome "medical reversal of the bottom line of a POEM", comprised fitting a binary logistic regression model to the data. Table 2 provides a summary of the estimated odds ratios and associated 95% confidence intervals from the logistic regression model.

Table 2. Odds ratio of predictive variables for the logistic regression model

	Odds Ratio	Confidence interval	
		2.5%	97.5%
Trial Sample Size	0.79	0.56	1.10
Sample Size Ratio	0.089	0.000 31	25.36
LOE Category	0.20	0.025	1.61
Trial Arms = 3	0.57	0.13	2.50
Trial Arms ≥ 4	1.00	0.17	5.93
Allocation Concealment = Uncertain	0.86	0.38	1.95
Drug / Non Drug	0.44	0.13	1.55
Year = 2003	6.51	1.38	30.67
Year = 2004	8.29	1.75	39.31
Year = 2005	9.32	1.89	45.90
Age Group	0.89	0.50	1.56
Supertype	1.50	0.57	3.93
Setting	1.19	0.77	1.85

In order to visualize the findings of the logistic regression in one comprehensive illustration, a nomogram was generated, allowing for identification of statistically and clinically relevant predictor variables (Figure 6).

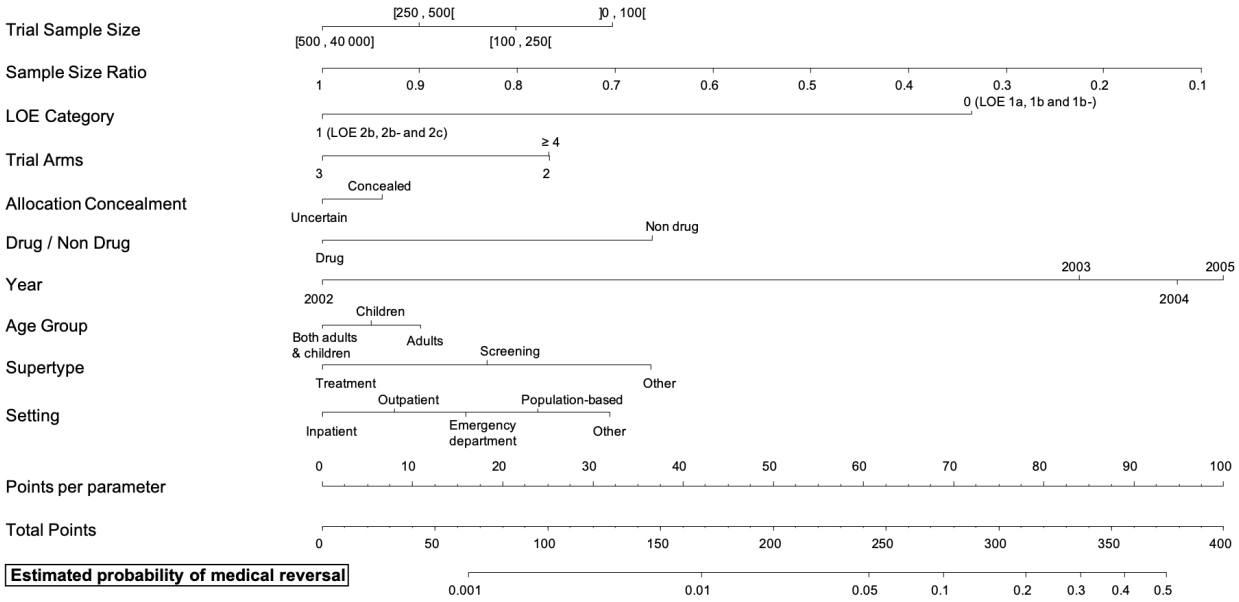

Figure 6. Nomogram #1 to predict the probability of the bottom line of a POEM to be reversed (an MR). All variables are included for Group 1 (reversed) and 2 (not reversed). Each variable is assigned a specific number of points. The sum of all those points corresponds to an estimate of the individual risk of MR.

Based on the nomogram, factors spanning a wider possible range of "Points" (upper scale) indicate a stronger association with the outcome of interest. Accordingly, the variables year, sample size ratio, LOE, supertype and whether the trial investigated a drug or not, displayed the strongest associations with the outcome of interest. In contrast, the factors 'allocation concealment' and 'age group' showed relatively weak associations with MR. In terms of the direction of the association, both the total sample size and the sample size ratio are negatively associated with MR. For the total sample size, this means conclusions from a larger trial have a lower chance to be reversed compared to findings from a smaller one. As for the sample size ratio, the smaller its value, meaning the smaller the intervention group size relative to the total sample size of the trial, the higher its probability of reversal. On the other hand, the year was positively associated with MR. In other words, more recent study findings were more likely to be reversed than findings from earlier studies.

The nomogram can be used to estimate the probability of the reversal of a POEM-RCT, given its individual characteristics. An example of how this estimation is calculated follows the second nomogram, presented below.

4.2.2.2 LASSO regression

As the prevalence of medical reversals was relatively low (n=34), including too many explanatory variables in the logistic regression model may have led to 'model overfitting' and therefore to possibly inaccurate parameter estimates (odds ratios and related nomogram subscales). In order to complement the conventional logistic regression analysis with a more robust estimation approach that accounts for overfitting, the results of the Least Absolute Shrinkage and Selection Operator (LASSO) regression model are displayed in Figure 7.

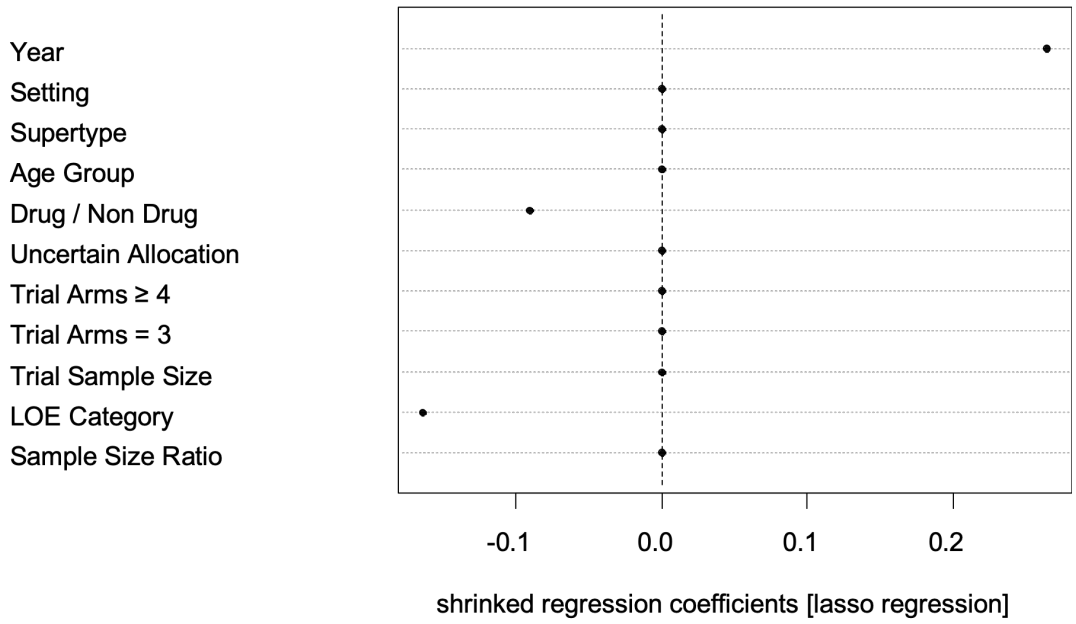

Figure 7. LASSO Regression Analysis

These results corroborate two findings of the conventional logistic regression analysis. First, the variables year, LOE and whether the trial investigated a drug or not were confirmed as being statistically associated with MR. Second, the directions of the associations of these three variables with the probability of an MR are the same as what is illustrated in nomogram #1.

In contrast to the conventional logistic regression analysis, the variable 'sample size ratio' was not strongly associated with MR. This means that the 'sample size ratio' may have less predictive capability than suggested by the likely overfitted logistic regression model.

Based on the findings of the LASSO regression, a second nomogram was created only including the confirmed predictive variables (Figure 8).

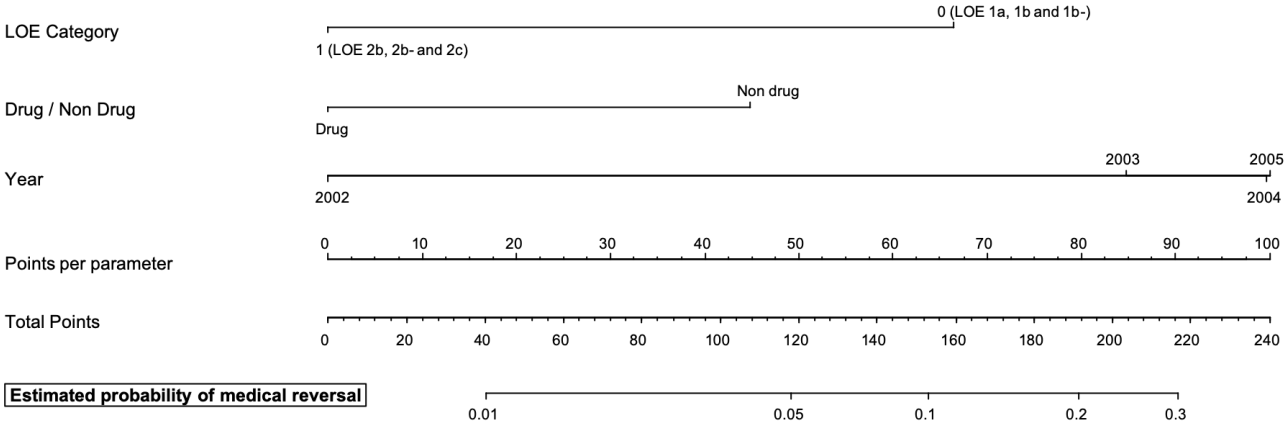

Figure 8. Nomogram #2 to predict the probability of the bottom line of a POEM to be reversed (an MR). Only the characteristics confirmed by the LASSO regression are included in this nomogram. Each characteristic is assigned a specific number of points. The sum of all points corresponds to an estimate of the individual risk of MR.

Nomogram #2 can be used to estimate the probability of a medical reversal of a specific POEM-RCT. I will illustrate this with a fictitious example: For a trial, investigating the use of an electronic monitoring device (45 score points), published in 2005 (100 score points) with a LOE of 1b- (~65 score points), the probability of MR would be of about 0.25 (total of 210 score points). However, if the same trial was published in 2003 (85 score points) and investigated a drug instead of a device (0 score points), then the probability of medical reversal would drop to just below 0.1 (total of 150 score points).

4.2.2.3 Classification Tree and Random Forest Analysis

With the objective of investigating the conditional differences in predictor variable levels across the groups, a classification tree and random forest were fitted.

The classification tree identified the variables year, setting, sample size ratio and LOE.cat as relevant variables determining statistical separation between studies that underwent MR and studies that did not (Figure 9). Each node of the tree displays the proportion of MRs in the

respective branch of the tree (centered value between 0 and 1 with corresponding color coding) and the percentage of individuals that fall in the respective variable subcategory or node (bottom % value).

According to the classification tree #1, the probability of reversal is its highest when the trial was not published in the year 2002, had an intervention group size to total sample size ratio lower than 0.5 and had a LOE of either 1a, 1b and 1b-. This echoes the three most relevant predictive variables identified by the conventional logistic regression and the lasso regression, apart from the sample size ratio.

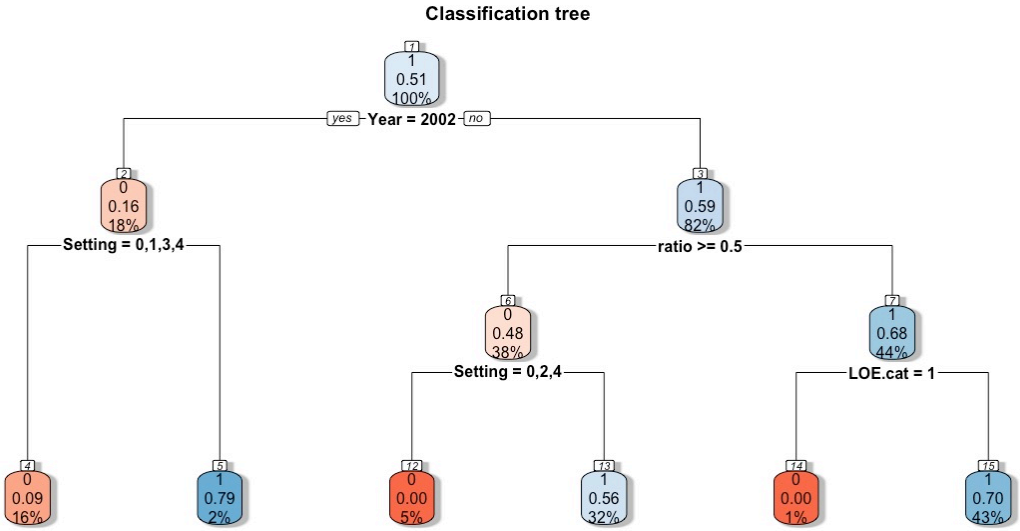

Figure 9. Classification tree #1 with all variables

The random forest analysis enabled computation of variable importance measures that reflect the decrease in the accuracy of the prediction model due removal of a respective candidate variable. The mean decrease in accuracy of prediction per variable removal is shown in Figure 10. Higher values indicate stronger predictive capability of a variable. It is important to note that the two generated random variables *randvar* and *randvar2* were correctly ranked to have zero explanatory capability with regard to predicting MR. In Figure 10, the variables year of publication, sample size ratio and total sample size were the three most important predictive variables according to the random forest analysis.

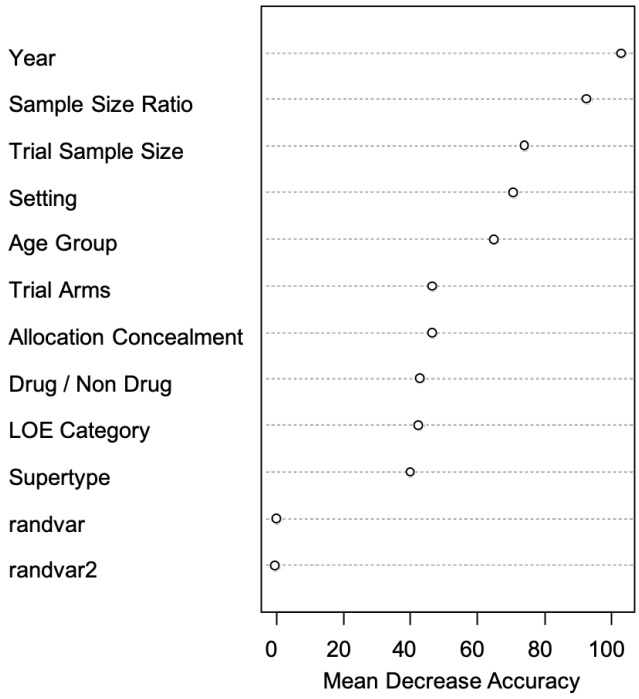

Figure 10. Random Forest #1 with all variables

In a subsequent analysis including only variables confirmed in the LASSO regression, the year and the LOE were still predictor variables but whether the trial investigated a drug or not does not seem to be very predictive (Figure 11).

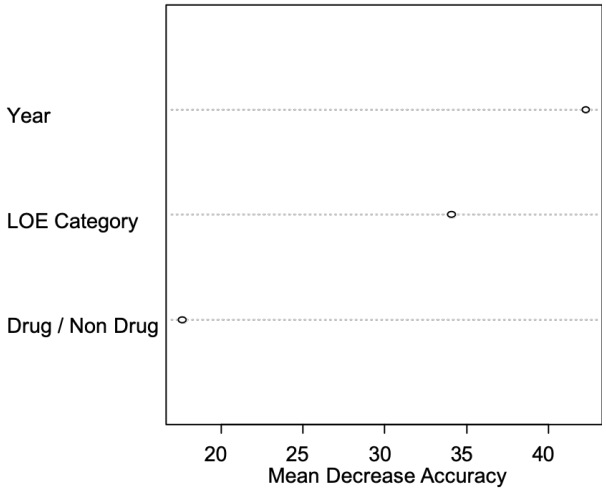

Figure 11. Random Forest #2 of the three variables confirmed by the LASSO regression

These findings are also reflected in the respective classification tree using the set of candidate variables identified through the LASSO regression (Figure 12). In this classification tree, whether the trial investigated a drug or not did indicate statistically robust predictive value.

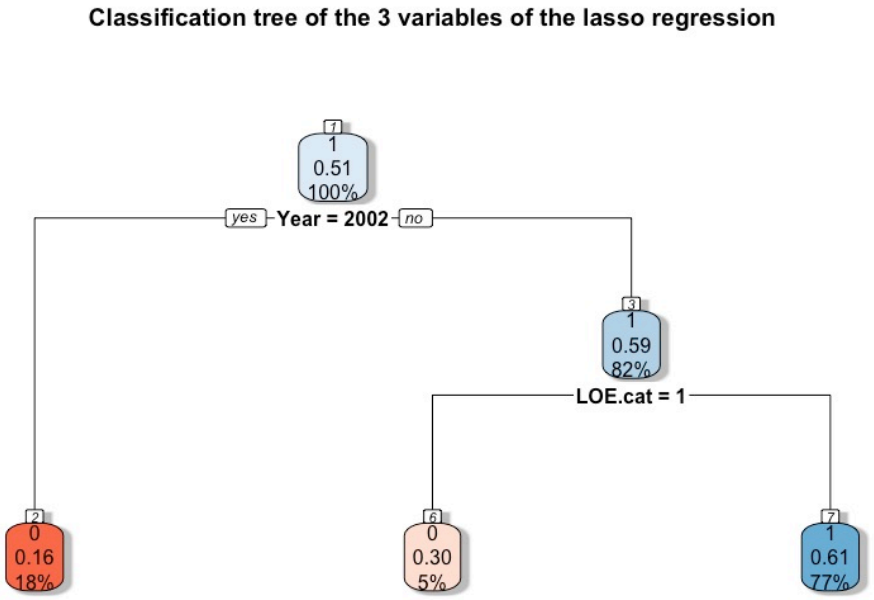

Figure 12. Classification tree #2 of the three variables confirmed by the LASSO regression

In summary, the random forests and classification trees confirm that year and LOE are statistically relevant predictor variables for MR. In addition, these analyses also confirmed the direction of associations found in the logistic and lasso regression analysis. For example, the LOE is negatively associated with MR, meaning that a POEM-RCT with a LOE of either 1a, 1b and 1b- has more chances of being reversed than one with 2b, 2b- and 2c. In the discussion section that follows, I will expand on these findings in the context of what is already known about the phenomenon of medical reversals.

V. Discussion

5.1 Background information

The objective of this study was twofold. First, I wanted to estimate how often POEM-RCTs become reversed. Prior research on this topic focused on the concept of medical practices becoming reversed by RCTs and was not specific to the context of primary care. Second, I sought to better understand the phenomenon of medical reversal by investigating whether the characteristics of trials could be associated with their reversal in the context of primary healthcare.

5.2 Key results

Of the 408 POEM-RCTs, we identified 34 (8.3%; 95% CI [6 , 11]) as being reversed in 2019. This represents a rate of reversal of 22 POEM-RCTs per 10 years. To my knowledge, this is the first estimation of reversals specifically for POEMs about RCTs.

As expected, this percentage of reversed POEM-RCTs is smaller than what has been reported in other work. In the context of internal medicine practice, Prasad's team conducted 3 studies. In study one: of 124 original research articles on medical interventions, 35 concerned an intervention that was already implemented by physicians. Of these, 16 (46%) were reversed (19). In study two: of 1344 original research articles on medical interventions, 146 concerned an intervention that was already implemented by physicians. Of these, 146 (40%) were reversed (15). In study two: of 3017 original research articles on medical interventions, 1644 concerned an intervention that was already implemented by physicians. Of these, 396 (24%) were reversed (44). we expected to find a lower percentage of reversals than Prasad for two reasons. First, Prasad studied the reversal of medical practices that were not necessarily supported by high quality RCT evidence. Second, we studied POEM-RCTs which theoretically should be less vulnerable to medical reversal than medical practices supported by studies that are based on lower level evidence (54). Indeed, more confidence can be placed on a medical practice if it is supported by an RCT than if supported by a retrospective cohort study, for example.

For the secondary objective concerning the association of study characteristics with POEM-RCT reversal, I conducted several modelling approaches. These involved a logistic and a lasso regression analysis, a classification tree, and a random forest, providing interesting findings. Only the year of study publication was identified as a strong predictive variable by all models. The

surprising aspect of this variable is in the direction of effect. Indeed, the results show a POEM-RCT published in 2005 has more chances of being reversed than one from 2002. This is counterintuitive since theoretically, the greater the time between E_1 and E_2, the more opportunity there is for a new study to be published on the same practice, potentially leading to a reversal. For this reason, this finding is probably due to chance, given the small number of events for the outcome of interest. Also, not all POEM-RCTs from 2005 were assessed. Due to time constraints, we assessed 85 of 92 POEM-RCTs published in 2005. This may have influenced these results. That being said, there may be another reason. Perhaps there is a positive association for the time between E_1 and E_2 but that this association plateaus after some years. In other words, there may be a positive association between year and reversal if we were looking at a population of trials that have been published less than 10 years prior. During that time, giving new evidence more time to emerge may be more significant than trials that are more than 10 years published, where this association may plateau. Regardless, this remains a testable hypothesis.

In addition, the LOE and the ratio were identified by three models each. In addition, the total sample size, the setting and whether or not the trial investigated a drug were identified by 1 model each. The findings relative to the total sample size suggests that larger trials have less chance of being reversed. This is because findings of smaller trials have smaller statistical power and thus an observed effect is more likely due to chance (125). The results also seem to suggest that the probability of reversal of a trial is diminished the higher is the proportion of the size of the intervention group to the total sample. This finding should be re-tested in future research.

As for the LOE, the results suggest that trials with LOE of 1a, 1b or 1b- have a higher chance of becoming reversed than trials with LOE of 2b, 2b- and 2c. Finding such an association is not surprising, but the direction of this association is counterintuitive. I expected most reversed POEMs to have a level of evidence of 2b, meaning "low quality randomized controlled trials (e.g. <80% follow-up)" (90), but there was only one such POEM with this LOE in the reversed group. In reality, 28 (80%) out of 34 reversed POEM-RCTs had a level of evidence of 1b, which stands for "individual randomized controlled trials (with narrow confidence interval)" (90). Based on these findings, in future work I recommend using a scale such as the Cochrane RoB. This will be further discussed below.

For the variable 'setting', only the classification tree suggested it could have an effect on MR, but when combined with other characteristics.

The last variable identified only by the lasso regression was whether the trial investigated a drug or not. This could be a spurious finding. Just 27 (6.6%) of the 408 POEM-RCTs did not investigate a drug and 4 of those were identified as reversed. As this finding is based on small numbers of outcome events, further research is needed.

On the topic of unanticipated findings, allocation concealment is one of mention. There was no unconcealed allocation in the subset of POEMs. This is unsurprising, as authors will rarely, if ever, admit to tampering with their trial. That being said, proper allocation concealment has been shown to be important to the integrity of randomization and in reducing the biases of any trial (70). Consequently, I was expecting uncertain allocation to be a shared characteristic among the reversed POEM-RCTs, which did not seem to be the case. In fact, the results suggest that concealed allocation is more strongly associated than uncertain allocation concealment, although the allocation concealment seems to be a weak predictive variable.

In *Figure 5. POEM-RCTs: Total Sample size per outcome*, there is one clear outlier in Group 1 (reversed). This POEM summarizes a subgroup of the Women's Health Study investigating the effect of low-dose aspirin on the risk of cancer in healthy women (126). This outlier embodies an important medical reversal as it introduces new parameters to the phenomenon, such as a variation in the baseline population risk. This notion is discussed in more detail later in this chapter. Another POEM was written with the same sample size based on the same trial, but this time investigating the effect of low-dose aspirin on alternate days on major cardiovascular events (127). This one, however was not reversed, hence the similar outlier in Group 2.

Finally, there is no mention of design and blinding in the results. This is because all 408 POEM-RCTs were double-blinded RCTs. Indeed, during data collection, all 960 POEMs were ordered by design and by year. The first design was double-blinded RCTs and we assessed all such trials for the years 2002-2004, and almost all for the year 2005.

5.3 Disagreements

When rating an MR, raters had to decide based on the supporting information I provided. Given their varying years of experience and knowledge, we did expect some disagreements. Since each rater has their own experience, some were more comfortable in rating POEM-RCTs. The discomfort lead some to further search the literature before making a recommendation. For

example, one particular rater was uncomfortable making a recommendation as s/he felt that a more exhaustive literature search was necessary to make such a rating. Even after consulting the literature, the reviewer had limited knowledge on the topic as her limited practice experience offered little or no exposure to the POEM topics. For this reason, the reviewer felt the recommendation should come from another reviewer to be more accurate. This situation demonstrates the uncertainty associated with the determination of MR. Moreover, this situation shows how different people will act based on what information is made available to them. From this example, it seems rather normal to observe a resistance to change practice after being reversed. Indeed, reversals can be controversial and difficult to accept.

In my study, 21 of the 34 (62%) POEM-RCTs identified as reversals were initially rated differently, or discordant between raters. This suggests that the concept of MR may not be as clear as we think. By now, it must be quite evident that the fluidity of evidence is taken into account when identifying MR. This variability makes the results of this research subject to change as new evidence arises. Thus, some of the POEM-RCTs identified as not reversed may be contradicted in future research. Furthermore, since identifying reversals can be a source of debate, it would not be surprising to have some readers disagree with our ratings.

I believe one of the main reasons explaining why MR remains a challenging concept comes from the distinction between MR and scientific progress. These are distinctly different situations. Let me explain further with a theoretical example. Imagine a drug called *F25*. This drug is used in practice to improve the bone density of postmenopausal women. While F25 can improve bone density, it comes with considerable side effects; yet the bone benefits are felt to outweigh the risks. If this theoretical drug were to exist, here are two situations to represent MR versus scientific progress: 1- A new and adequately powered trial reveals that *F25* increases the risk of cardiovascular disease, thus any benefits are now outweighed by a discovery of new and more important harms; 2- Along comes a new drug (*F25+)* offering the same benefits, but with drastically lower risks of harm. In situation 1, the drug is shown to be more harmful than originally claimed. The emergence of this new evidence on F25 would lead clinicians to a MR as their change in prescribing practice would be to stop it. In situation 2, I describe a scenario where a better drug emerges. The effectiveness of F25 and associated risk of harms are not modified in any way by the appearance of *F25+*. But the practice change here would be to prescribe *F25+* instead of *F25*. This represents an example of scientific progress.

5.4 Type of reversal

Of the 34 POEM-RCTs identified as reversed, 17 (50%) were found to have a change in the direction of effect from positive to negative. The other half was from negative to positive. This is quite surprising. First, we found as many changes of direction of effect from positive to negative as from negative to positive. In another study where changes in the direction of effect were reported, the situation where the original study reports efficacy of a treatment (positive) while the later study reports a lack of efficacy or harms (negative) happened almost four times more often (n= 14, 58% vs n= 4, 16%) (29). Second, I would be curious to further investigate the rationale for a trial of a therapy that has previously been reported to be ineffective or harmful. I can understand how therapies with sparse evidence supporting their use may receive ethical approval, however it strikes me as odd that a new trial of a therapy can be initiated when prior evidence shows either no efficacy or harm (128). Further investigation is necessary to understand the reasons for reversal, when the direction of effect shifts from negative to positive.

From the literature, a shift in the direction of effect from negative to positive could be due to the wrong population being studied or an inappropriate dosing regimen (129). One way of minimizing the latter is to have multiple trial arms with different dosage of the treatment regimen. For example, a trial investigating migraine prevention compared placebo to increasing doses of topiramate (50mg/d; 100mg/d; 200mg/d) (130).

Regardless, I believe it is important to broaden our existing definition of MR. My proposal for a revised definition is as follows: an MR occurs when the direction of treatment effect changes due to new and better-quality evidence; and this contributes to practice change, regardless of whether the initial claim was in support of or against a given medical practice. In this way, an MR will not be limited to a specific change in the direction of clinical practice.

5.5 Limitations

The most significant limitation of this work is the small number of reversed POEM-RCTs we found. The frequency of the outcome of interest particularly affects the estimates of the strength of association between trial study characteristics and a reversal. Indeed, although the statistical modelling approaches used in this work are very powerful to find predictive variables, the small number of reversals, these approaches were used as descriptive rather than predictive tool. Thus, with a bigger sample size, more reliable information could have been generated. For now, this

work remains exploratory and hypothesis generating. Since this limited number of medical reversals may be shared to inspire others, I plan on sharing my dataset on an online open-access platform to support further research.

Having an estimate of the proportion of reversals among POEM-RCTs over time gives no indication as to how often these medical practices are used in primary care. Furthermore, as mentioned before, the screening process applied to the primary literature for the selection of studies to be summarized as POEMs filters all research for specific criteria. Although it may be difficult to know how this step impacted our study, we believe our findings in Step 1 underestimate the rate of reversal in all RCTs.

Though previously mentioned, the identification of MR is influenced by time. Since evidence is fluid, reproducing the study in a few years with the same subset of POEM-RCTs may result in different estimates of MR. On that note, we did not differentiate between POEMs that were identified as not reversed (when new evidence confirms prior findings) and others that had no new evidence published since the POEM in Step 1. Therefore, new reversed POEM-RCTs will be identified over time as new evidence emerges.

This study evaluated various trial characteristics as potential predictors of MR. In other words, other sources of MR were not evaluated. Probably the most important is the influence of innovation in science and technology (external factors that affect research studies unrelated to trial design per se). For example, when a new method of analysis emerges and helps to improve the data, then the results of all subsequent studies may be affected. This effect might be solely due to this new method.

Another limitation is that the extracted data can only be as good as the quality of the reporting in the RCT. In addition to poor reporting, there is the potential for human error in the coding of variables such as allocation concealment within the POEMs database. That being said, as previously mentioned in the results section I found just 3 of the 410 POEM-RCTs (0.7%) to be misclassified as double-blinded RCT. Thus, the effect of this limitation is believed to be very minimal.

Finally, we did not differentiate between a reversal of a biased RCT and one of a well-conducted RCT where a technological advance, such as a better alternative therapy, affected the recommendation. In other words, whether the reversal is due to the characteristics of the study and how it was conducted or due to external factors. For example, aspirin was recommended for

the primary prevention of cardiovascular disease for decades. Due to a decline in the population risk for cardiovascular disease such as a reduced prevalence of smoking, the use of aspirin as primary prevention has been reversed; the gastrointestinal risk of harm in 2019 is now perceived to equal or outweigh any cardiovascular benefits (64). This reversal was also the result of having a new preventive therapy, namely statins.

5.6 What would I have done given more time?

In an ideal world, I would have recruited more raters and rated all 910 POEM-RCTs. As a master's thesis, this project was limited in terms of time and human resources. Data collection ran from September 2018 until the end of February 2019. During that time, I had to prepare the coding template for the raters, compile the agreements and disagreements, and extract the study characteristics from each POEM-RCT. It also took time for a physician-rater to familiarize themselves with the concepts and the coding process. A major limiting factor was the process of data collection in Step 1, for two reasons. First, preparing the coding template – searching and extracting E_2 from DynaMed – was time consuming. Indeed, some topics were very easy to access but many would take more than 15 minutes to extract. This was dealt with by adding a 15-minute time limit for each topic. Second, the amount of work demanded from the raters was also a challenge. As mentioned before, raters received sets of 50 POEM-RCTs to code at one time. Such sets would take anywhere between 1 to 5 hours to code. This was volunteer work added to an already heavy work load. Afterwards, I had to identify any disagreements, a step I could only do once I received ratings of both raters. Then, the two raters needed to meet to discuss their decisions. When some disagreements persisted after a single discussion, they would be sent to a third rater, further lengthening the process. In most cases, the disagreements were resolved by the pair of raters, without the need for a 3rd rater. Only after the codes were obtained for the whole set of 50 POEMs would I ask if they were available for another set.

If I had more time, I would have extracted more study characteristics. Indeed, I would have devised a clear operational definition and extracted the duration of the study, the industry sponsorship, the field of research, the type of analysis, the medical specialty associated with the topic of the trial, the effect size, the Risk of Bias score, the different levels of blinding, whether

or not the trial was registered, terminated early, multi-centered, in addition to the authors' affiliations for potential conflict of interest.

The duration of the study and early termination of the trial are interesting variables as short trials may not offer an appropriate representation of the effect of the intervention. Indeed, research has shown that trials stopped early with a beneficial effect tend to be shown to have a larger estimated effect size as compared to subsequent trials (131, 132). There are also concerns of misinterpretation of effect when a trial is terminated early with the intervention showing clear benefits (133). For the breakdown of industry sponsorship, I would extract two components: the funding source (industry, non-industry, combination of industry and non-industry, etc.) and whether or not the authors' affiliations are potential sources of conflict (yes, no, uncertain) (44). As for the field of research, in some cases, it is very straightforward and thus easy to extract. For example, a POEM summarizes the PEACE trial that investigated whether adding an angiotensin-converting enzyme inhibitor improves outcomes in patients with stable angina and no evidence of heart failure (134). In that case, it is clear that the field of study is cardiology. It is more challenging when there is a disease that affects multiple organs or when a study investigates the effect of a drug on cardiac symptoms, but its pathophysiology affects the lungs, for example. In the case of single- or multi-center study, a single center can drive an intervention to be beneficial, however, the effect seen may be associated to the center itself and not the intervention (135).

I would also like to add the effect size as a variable. As mentioned in my literature review, although small effect size may be erroneous in future study (125), large effect sizes may be overestimates caused by biases (75-77). It would be interesting to examine the relationship between effect size and MR.

As seen in the results, the LOE was slightly associated with reversal and the direction of this association was counterintuitive. Indeed, my results seem to suggest that reversal is more associated with higher LOE, better quality, than with lower LOE. For this reason, I would consider using Cochrane's RoB tool in future work given that it may have greater validity in the measurement of trial quality. This is mainly because the LOE scale seems more subjective than the RoB tool in its assessment of trial quality. Indeed, each level of the scale seems to be self-explanatory, however there is a lack of clarity regarding the assignment of levels to each

POEMs-RCT (90). When looking at the LOE scale, an RCT should only be able to fit under the following levels:

- 1b: Individual randomized controlled trials (with narrow confidence interval);
- 1c: All or none randomized controlled trials;
- 2b: Individual cohort study or low quality randomized controlled trials (e.g. <80% follow-up).

In addition, if a minus sign "-" is added to any of the above levels, it means that the evidence "fails to provide a conclusive answer because it is (…) a single result with a wide Confidence Interval" (136).

In other words, the only criteria that seem to be evaluated, or at least the only ones mentioned in the very brief definition in the scale of levels, concern whether the trial was well conducted or not and the loss to follow-up. It remains uncertain whether more criteria are taken into consideration when an LOE is assigned. On another note, in the 408 data entry studied, some POEM-RCTs were assigned levels 1a and 2c, which does not seem to possible given the RCT designs; granted these may be due to human error. As for Cochrane's tool, the assigned RoB seems to be more objective. First, the process of assigning a risk of bias by using Cochrane's RoB tool involves the independent assessment of each study by two reviewers, with any disagreements resolved by a third reviewer (source). Second, more elements are accounted for in the RoB tool than in the LOE. In fact, the RoB tool assesses seven sources of biases, including but not limited to allocation concealment, blinding of participants, of personnel, and of outcome assessment. Third, the Cochrane RoB tool is specific to trials. For these reasons, I would recommend the RoB tool for a more valid scale assessing trials.

The analysis used for the trial would also be another data point. Indeed, the type of analysis used can have an impact on the results as it can affect the randomization and introduce some biases (137). This could be extracted as the analysis being intention-to-treat (ITT), per-protocol (PP) and uncertain. Trial registration could also be added. One of the reasons for registering a trial is to reduce selective reporting (138), publication bias and analytical flexibility, such as outcome switching. As a small proportion of registered trials publish outcomes completely consistent with their pre-registration plan (34), whether or not a trial has been pre-registered would also be of interest.

In addition, I would extract not only how many blinding levels but also who was blinded. Indeed, all double-blinded randomized trials do not all blind the participant and the health care provider. For this reason, instead of simply categorizing RCTs by design (single-blinded, double-blinded, etc.) information on blinding would be extracted under 4 levels: participants (subjects), health care providers, data collectors and the people in charge of outcome assessment (71). As many biases are reduced due to masking, future studies should investigate if any difference in blinding of trials is associated with the occurrence of MR. My hypothesis is that rates of reversals will be higher in single-blinded trials compared to those double-blinded, while non-blinded trials will have the highest rate of MR.

5.7 Lessons learned from the methodology

One aspect of the method that spurred many discussions with my supervisor and other contributors was whether one source of evidence was enough to identify MRs. Underlying this discussion is the notion of how much certainty is needed before making a decision. In this study, we used DynaMed because this information source is known to have a strong surveillance system in place to scan the literature and include new evidence. This system, however, is not necessarily perfect and a thorough PubMed search may have yielded different results. That being said, doing such an extended search for each POEM-RCT not only demands a lot of work, it also does not represent what healthcare professionals are able to do at the point of care or even to keep updated on a topic (82). Many healthcare professionals that search for answers will use their favorite knowledge resource, for example DynaMed. Since we wanted to recreate what healthcare professionals would do, we focused on a single knowledge resource.

As mentioned before, raters were free to further search the literature if they felt it was necessary. At least one team of raters searched additional knowledge resources in Step 1, especially when they disagreed on the coding of a particular POEM-RCT. Out of a set of 50 POEM-RCTs, 3 were coded based on a different source of evidence than DynaMed. Although none of the 3 were found to be reversed, it is also possible that using a second knowledge resource could have impacted my results. For this reason, I would advise future researchers to explore the possibility of using 2 resources, for example DynaMed and UpToDate. Indeed, even if knowledge resources have their own literature surveillance systems, remaining updated is a difficult task (3).

5.8 Is there a higher rate of MR in POEMs Observational Studies?

As an ad hoc exploration, I wanted to compare the rate of reversal in POEMs of studies other than RCTs. As observational studies are lower on the pyramid of evidence when it comes to questions of therapy, they could be associated with a higher rate of MR (54).

For this exploratory work, I selected POEMs summarizing case-control, prospective and retrospective cohort studies from the POEMs database provided by Dr. Ebell. These POEMs were assigned a random number and then sorted in ascending order. Subsequently, my supervisor and I read 100 POEMs one by one, in a sequential manner, to select only the POEMs and bottom lines that could potentially be reversed over time. For example, POEMs about the predictors of survival after a diagnosis of Alzheimer's disease or the prognosis of different kinds of syncope have less chance of being reversed, so we did not include them in our selection. We reviewed the first 100 POEMs to select 51 where a reversal was deemed to be possible. With this subset of 51 POEMs, we applied the process described in Step 1 for the identification of reversed POEMs. In other words, the bottom line of the POEM and the updated evidence was examined by two independent reviewers to identify the frequency of MR, with a third reviewer in case of disagreement.

Of 51 POEM-observational studies (POEM-Obs), 37 (72.5%) are from prospective cohort, 8 (15.7%) are from retrospective cohort, and 6 (11.8%) are from case-control studies. The pair of raters disagreed on 15 occasions. 5 of those were reconciled through discussion. A 3rd physician-rater resolved the other 10 disagreements.

From this random sample of 51 POEM-Obs, 12 (23.5% ; 95% CI [14 , 36.8]) were identified as reversed. As one would expect, the phenomenon of MR occurs more frequently within POEM-Obs (23.5%) than POEM-RCT (8.3%), with no overlap of the confidence intervals (95% CI [14 , 36.8] vs [6 , 11] respectively). This result is comparable to what has been reported in the literature, where three studies identified 11-13% of original articles concerning a medical practice and 24-46% of original studies on already adopted medical practices as being reversals (15, 19, 44). Perhaps our selection process could have influenced the number of reversed POEMs found, but since the result is comparable to that reported, the sturdiness of our method is confirmed.

5.9 Should evidence come with an expiration date?

A very interesting paper discussed the notion that evidence may need an expiration date. This idea emerged from a finding that therapies such as ASA, previously proven to be effective, can 'stop working' (61). Reasons for this phenomenon include changes in population risk for disease and the use of new and better therapies (139). With this in mind, future studies may want to integrate more information about the population at risk and their treatment. For example, in the case of cardiovascular disease prevention, is gender represented in equal proportion? What is the prevalence of risk factors such as diet, exercise and smoking, in the population? These new parameters may not explain the primary cause of a reversal, but they may reveal something of value in trying to understand how and why a reversal has occurred.

This paper is of great importance as it shows that external factors may help to explain a reversal. This is interesting as it questions the potential validity of any predictive model for MR. More precisely, a predictive model will only be useful if external factors play a smaller role in the phenomenon of reversal than the study characteristics themselves.

5.10 Future Implications

A rate of reversal of 22 POEM-RCTs per 10 years carries several implications. First, EBM educators need to be aware of this phenomenon; about 22 medical practices supported by RCTs will be shown to be ineffective or harmful in 10 years' time. For practice, the consequence of these changes is unknown as we have ignored whether these practices are very common in primary care. In addition, physicians using POEMs as a source for updating their knowledge will eventually read a synopsis contradicting what they do. Knowing about this phenomenon will help to better manage the experience of conflicting evidence. Health administrators are also affected by the rate of MR as policies to cover therapies will need to be reviewed. Furthermore, the editors of knowledge resources are affected, as they must offer updated and valid evidence. Finally, there are implications for patients because of the direct negative impact on their health from ineffective or harmful therapy.

For all of these reasons, awareness of this phenomenon is a first step. In this regard, publishing lists of reversed practices with a rationale as to why the practice should be changed (15) is a great start, but the impact of this list depends on its uptake into practice. As a complement, I

envision a two-part flagging or tagging system of reversed medical practices put into place by high-quality resources of updated information, for example EE+. The first part would be to flag the practice itself with a grading system. For convenience sake, this could be added to a system such as the Grading of Recommendations Assessment, Development and Evaluation (GRADE) (140). GRADE is a 4-level system for rating the quality of evidence and the strength of recommendations. The different levels of quality of evidence in this system acknowledge that low quality evidence is likely to be modified as new evidence emerges, but there is no mention of a change in practice following a reversal. An additional grade, for example "R", could be assigned to a practice that has been reversed. This would help to inform people about any practice that is no longer recommended. The 34 POEM-RCTs that we identified as reversed should be labelled as 'R'. In the future, my results suggest an 'R' grade will need to be assigned to POEM-RCTs at a rate of about 2 per year. This does not consider the rate of reversal in POEMs of other study design. Nevertheless, this work could allow the EE+ editors to further improve the POEMs database.

Part 2 would be to provide a supplementary piece of information, which would be a brief explanation of why a medical practice was reversed, with appropriate references. The rationale for this supporting explanation is that it will help stakeholders to understand the reasons behind this label of reversal. Hopefully, this process can assist the integration of new evidence into clinical practice. Examples of these explanations can be seen in the supplemental documents in Prasad et al (15, 44). In Prasad et al. 2013, in the column "How it contradicted existing medical practice?" an explanation reads as follows: "Many patients with persistent symptoms of Lyme disease receive prolonged courses of antibiotics, although the effectiveness of this practice remains unknown[13]. This randomized, placebo-controlled, double-blinded trial failed to show any significant improvement in symptoms after a prolonged 90-day course of antibiotics in patients with persistent symptoms."

As discussed throughout this thesis, medical evidence is always subject to a specific context; a time and a place. Indeed, even practices supported by robust evidence have been reversed (61). Thus, maintaining correct information in any knowledge resource requires continuous updating. However, such maintenance brings forth many considerations, namely the cost, assigning responsibility for the task, and how often? Of course, the updating of knowledge resources challenges human resources.

As for the association between study characteristics and reversed POEM-RCTs, work is needed to confirm my findings. In an ideal world, with a bigger sample size, perhaps an algorithm could be generated. This predictive algorithm could help healthcare professionals learn the probability of reversal of any one RCT. Such a tool may help in the decision-making process. For example, in a situation where a physician reads about a new trial, knowing the probability of reversal may influence a prescribing decision.

Until then, I contend that awareness is the first step to confront the phenomenon of MR.

VI. Conclusion

The main goal of this research was to better understand the phenomenon of medical reversal in primary care, with two objectives. The first objective was to reveal the rate of medical reversal in POEM-RCTs. The second objective was to explore whether any characteristics of POEM-RCTs are associated with MR in primary healthcare.

An initial sample of 960 POEM-RCTs from 2002-2007 was studied in a two-step process. In step 1, pairs of raters independently categorized POEM-RCTs as reversed (or not) by new evidence. In step 2, characteristics of POEM-RCTs were extracted as factors, to investigate the association between these factors and the outcome of interest.

After completing these two steps, 408 POEM-RCTs were analyzed. Using descriptive statistics and the modelling approaches of multiple logistic regression, Least Absolute Shrinkage and Selection Operator (LASSO) regression, classification trees, and random forest, the relationship between the outcome of interest and these study characteristics was investigated.

My first objective was met. Indeed, with the help of raters, 34 POEM-RCTs were identified as reversed, corresponding to a rate of about 22 POEM-RCTs per 10 years. In other words, 2 POEM-RCTs will be reversed per year. As mentioned before, this remains an approximation and I believe this rate is an underestimation of the true number of reversals in the medical literature.

As for my second objective, associations between characteristics of trials and the outcome of interest were identified. Indeed, the year, the level of evidence (LOE) and the ratio emerged as potentially predictive variables. However, since the number of MRs was modest, further investigation with a larger sample size is necessary.

The findings of this research have implications for many stakeholders. The editors of knowledge resources already work to update their resources. Flagging of POEMs that have been identified as reversed should be considered as an updating task. In addition, physicians who use the daily POEMs need to be aware of this phenomenon as they will encounter new evidence that contradicts their current practice. In addition, awareness of the phenomenon will help physicians better manage the experience of conflicting evidence. In the same vein, teachers of evidence-

based medicine should update their curricula to increase awareness of the phenomenon. All those implications are important to reduce the harms associated with a medical reversal.

A future avenue of research is the development and testing of a model to predict the probability of medical reversal of original clinical research. Such a tool could be helpful in improving care delivery and in medical education. Indeed, if a clinician knew the probability of reversal associated with any single RCT, then s/he could consider this issue in a shared decision-making context. Knowing the probability of reversal may also help to delay the use of a practice until confirmatory and satisfactory evidence arises.

Finally, being able to better predict medical reversal does not mean we will be able to avoid the phenomenon. Indeed, an MR can occur due to external factors, such as a change in the baseline population risk for a disease (61). For this reason, although we should strive to minimize the impact of MR before it happens, we also need to be able to manage the reaction to an MR once it has been identified. The de-implementation of a practice shown to be ineffective or harmful remains a separate challenge.

References

1. Sackett DL, Rosenberg WM, Gray JA, Haynes RB, Richardson WS. Evidence based medicine: what it is and what it isn't. British Medical Journal. 1996;312(7023):71-2.
2. Pierre Pluye, Roland Grad, Julie Barlow. Look It Up! What Patients, Doctors, Nurses, and Pharmacists Need to Know about the Internet and Primary Health Care: McGill-Queen's University Press; 2017 October 2017.
3. Bastian H, Glasziou P, Chalmers I. Seventy-five trials and eleven systematic reviews a day: how will we ever keep up? PLoS Med. 2010;7(9):e1000326.
4. Ioannidis JP. Evidence-based medicine has been hijacked: a report to David Sackett. J Clin Epidemiol. 2016;73:82-6.
5. Cosgrove L, Vannoy S, Mintzes B, Shaughnessy AF. Under the Influence: The Interplay among Industry, Publishing, and Drug Regulation. Account Res. 2016;23(5):257-79.
6. Lexchin J, Bero LA, Djulbegovic B, Clark O. Pharmaceutical industry sponsorship and research outcome and quality: systematic review. BMJ (Clinical research ed). 2003;326(7400):1167.
7. Every-Palmer S, Howick J. How evidence-based medicine is failing due to biased trials and selective publication. J Eval Clin Pract. 2014;20(6):908-14.
8. Lewis SJ, Orland BI. The importance and impact of evidence-based medicine. J Manag Care Pharm. 2004;10(5 Suppl A):S3-5.
9. Ioannidis JPA, Stuart ME, Brownlee S, Strite SA. How to survive the medical misinformation mess. Eur J Clin Invest. 2017;47(11):795-802.
10. Smith AB, Semler L, Rehman EA, Haddad ZG, Ahmadzadeh KL, Crellin SJ, et al. A cross-sectional study of medical student knowledge of evidence-based medicine as measured by the Fresno test of evidence-based medicine. The Journal of emergency medicine. 2016;50(5):759-64.
11. Prasad VK, Cifu AS. Ending Medical Reversal: Improving Outcomes, Saving Lives. Baltimore: Johns Hopkins University Press; 2015.
12. Sutton D, Qureshi R, Martin J. Evidence reversal - when new evidence contradicts current claims: a systematic overview review of definitions and terms. J Clin Epidemiol. 2018;94:76-84.
13. Manson JE, Hsia J, Johnson KC, Rossouw JE, Assaf AR, Lasser NL, et al. Estrogen plus progestin and the risk of coronary heart disease. New England Journal of Medicine. 2003;349(6):523-34.
14. Boardman HM, Hartley L, Eisinga A, Main C, Roqué i Figuls M, Bonfill Cosp X, et al. Hormone therapy for preventing cardiovascular disease in post-menopausal women. Cochrane Database of Systematic Reviews. 2015(3).
15. Prasad V, Vandross A, Toomey C. A decade of reversal: an analysis of 146 contradicted medical practices. Mayo Clinic proceedings. 2013;88.
16. Prasad V, Cifu A, Ioannidis JA. Reversals of established medical practices: Evidence to abandon ship. JAMA. 2012;307(1):37-8.
17. Cifu AS, Prasad VK. Medical Debates and Medical Reversal. Journal of General Internal Medicine. 2015;30(12):1729-30.

18. Antman EM, Lau J, Kupelnick B, Mosteller F, Chalmers TC. A Comparison of Results of Meta-analyses of Randomized Control Trials and Recommendations of Clinical Experts: Treatments for Myocardial Infarction. JAMA. 1992;268(2):240-8.
19. Prasad V, Gall V, Cifu A. The frequency of medical reversal. Archives of internal medicine. 2011;171(18):1675-6.
20. Zhang W. Meta-Epidemiology: Building the Bridge From Research Evidence to Clinical Practice. Osteoarthritis Cartilage. 2010;18:S1.
21. Murad MH, Wang Z. Guidelines for reporting meta-epidemiological methodology research. Evid Based Med. 2017;22(4):139-42.
22. Phillips DC, Burbules NC. Postpositivism and educational research. Lanham, Md.: Rowman & Littlefield Publishers; 2000.
23. Hong QN, Pluye P, Bujold M, Wassef M. Convergent and sequential synthesis designs: implications for conducting and reporting systematic reviews of qualitative and quantitative evidence. Syst Rev. 2017;6(1):61.
24. Creswell JW, Creswell JD. Research design : qualitative, quantitative, and mixed methods approaches. Fifth edition. ed. Thousand Oaks, California: SAGE Publications, Inc.; 2018.
25. Gough D, Oliver S, Thomas J. An introduction to systematic reviews: Sage; 2017.
26. Ilic D, Djulbegovic M, Jung JH, Hwang EC, Zhou Q, Cleves A, et al. Prostate cancer screening with prostate-specific antigen (PSA) test: a systematic review and meta-analysis. BMJ (Clinical research ed). 2018;362:k3519.
27. Abad C, Fearday A, Safdar N. Adverse effects of isolation in hospitalised patients: a systematic review. J Hosp Infect. 2010;76(2):97-102.
28. Ruchon C, Grad R. Evidence reversal: Towards awareness of the phenomenon in library and information science. Education for Information. 2018(Preprint):1-4.
29. Niven DJ, McCormick TJ, Straus SE, Hemmelgarn BR, Jeffs L, Barnes TRM, et al. Reproducibility of clinical research in critical care: a scoping review. BMC Medicine. 2018;16(1):26.
30. Gnjidic D, Elshaug AG. De-adoption and its 43 related terms: harmonizing low-value care terminology. BMC Medicine. 2015;13(1):273.
31. Bollen K, Cacioppo JT, Kaplan RM, Krosnick JA, Olds JL, Dean H. Social, behavioral, and economic sciences perspectives on robust and reliable science. Report of the Subcommittee on Replicability in Science Advisory Committee to the National Science Foundation Directorate for Social, Behavioral, and Economic Sciences. 2015;3:3.
32. Lawton JS. Reproducibility and replicability of science and thoracic surgery. J Thorac Cardiovasc Surg. 2016;152(6):1489-91.
33. Lanspa MJ, Hirshberg EL, Miller RR, 3rd, Morris AH. Clinical study replicability and the pursuit of excellence. Crit Care. 2015;19:297.
34. Munafò MR, Nosek BA, Bishop DVM, Button KS, Chambers CD, Percie du Sert N, et al. A manifesto for reproducible science. Nature Human Behaviour. 2017;1:0021.
35. Makel MC, Plucker JA, Hegarty B. Replications in Psychology Research: How Often Do They Really Occur? Perspect Psychol Sci. 2012;7(6):537-42.
36. Ioannidis JPA. Acknowledging and Overcoming Nonreproducibility in Basic and Preclinical Research. JAMA. 2017;317(10):1019-20.

37. Baker M. 1,500 scientists lift the lid on reproducibility. Nature News. 2016;533(7604):452.
38. Carney DR, Cuddy AJC, Yap AJ. Power Posing: Brief Nonverbal Displays Affect Neuroendocrine Levels and Risk Tolerance. Psychol Sci. 2010;21(10):1363-8.
39. Ranehill E, Dreber A, Johannesson M, Leiberg S, Sul S, Weber RA. Assessing the Robustness of Power Posing: No Effect on Hormones and Risk Tolerance in a Large Sample of Men and Women. Psychol Sci. 2015;26(5):653-6.
40. Simmons JP, Simonsohn U. Power Posing: P-Curving the Evidence. Psychol Sci. 2017;28(5):687-93.
41. Cuddy AJC, Schultz SJ, Fosse NE. P-Curving a More Comprehensive Body of Research on Postural Feedback Reveals Clear Evidential Value for Power-Posing Effects: Reply to Simmons and Simonsohn (2017). Psychol Sci. 2018;29(4):656-66.
42. Goodman SN, Fanelli D, Ioannidis JP. What does research reproducibility mean? Sci Transl Med. 2016;8(341):341ps12.
43. Ebrahim S, Sohani ZN, Montoya L, Agarwal A, Thorlund K, Mills EJ, et al. Reanalyses of Randomized Clinical Trial DataReanalyses of Randomized Clinical Trial DataReanalyses of Randomized Clinical Trial Data. JAMA. 2014;312(10):1024-32.
44. Herrera-Perez D, Haslam A, Crain T, Gill J, Livingston C, Kaestner V, et al. A comprehensive review of randomized clinical trials in three medical journals reveals 396 medical reversals. eLife. 2019;8:e45183.
45. Ioannidis JP. Contradicted and initially stronger effects in highly cited clinical research. Jama. 2005;294(2):218-28.
46. Schpero WL. Limiting low-value care by "choosing wisely". Virtual Mentor. 2014;16(2):131-4.
47. Schpero WL. Limiting low-value care by “choosing wisely”. AMA Journal of Ethics. 2014;16(2):131-4.
48. Elshaug AG, Watt AM, Mundy L, Willis CD. Over 150 potentially low-value health care practices: an Australian study. The Medical journal of Australia. 2012;197.
49. Guyatt GH, Oxman AD, Kunz R, Vist GE, Falck-Ytter Y, Schünemann HJ. What is “quality of evidence” and why is it important to clinicians? BMJ (Clinical research ed). 2008;336(7651):995-8.
50. Frakt A. 2013, January 16. [cited 2019]. Available from: https://theincidentaleconomist.com/wordpress/half-of-medical-treatments-of-unknown-effectiveness/.
51. Hanson R. 2009, July 31. Available from: http://www.overcomingbias.com/2009/07/meds-to-cut.html.
52. Prasad V, Jorgenson J, Ioannidis JP, Cifu A. Observational studies often make clinical practice recommendations: an empirical evaluation of authors' attitudes. J Clin Epidemiol. 2013;66(4):361-6.e4.
53. Mellis C. Evidence-based medicine: What has happened in the past 50 years? J Paediatr Child Health. 2015;51(1):65-8.
54. Murad MH, Asi N, Alsawas M, Alahdab F. New evidence pyramid. Evidence Based Medicine. 2016.

55. Prasad V. Plenary Session [a podcast on medicine, oncology, & health policy] [Internet]; 2018. Podcast. Available from: https://soundcloud.com/plenarysession/127-bonus-medical-reversal
56. Khan KS, Kunz R, Kleijnen J, Antes G. Five steps to conducting a systematic review. J R Soc Med. 2003;96(3):118-21.
57. Westfall JM, Mold J, Fagnan L. Practice-based research—"blue highways" on the nih roadmap. JAMA. 2007;297(4):403-6.
58. Burns PB, Rohrich RJ, Chung KC. The levels of evidence and their role in evidence-based medicine. Plast Reconstr Surg. 2011;128(1):305-10.
59. Hariton E, Locascio JJ. Randomised controlled trials - the gold standard for effectiveness research: Study design: randomised controlled trials. BJOG : an international journal of obstetrics and gynaecology. 2018;125(13):1716-.
60. Higgins JPT, Altman DG, Gøtzsche PC, Jüni P, Moher D, Oxman AD, et al. The Cochrane Collaboration's tool for assessing risk of bias in randomised trials. BMJ (Clinical research ed). 2011;343.
61. Greene P, Prasad V, Cifu A. Should Evidence Come with an Expiration Date? Journal of General Internal Medicine. 2019.
62. Final Report on the Aspirin Component of the Ongoing Physicians' Health Study. New England Journal of Medicine. 1989;321(3):129-35.
63. Sanmuganathan PS, Ghahramani P, Jackson PR, Wallis EJ, Ramsay LE. Aspirin for primary prevention of coronary heart disease: safety and absolute benefit related to coronary risk derived from meta-analysis of randomised trials. Heart. 2001;85(3):265-71.
64. Gaziano JM, Brotons C, Coppolecchia R, Cricelli C, Darius H, Gorelick PB, et al. Use of aspirin to reduce risk of initial vascular events in patients at moderate risk of cardiovascular disease (ARRIVE): a randomised, double-blind, placebo-controlled trial. The Lancet. 2018;392(10152):1036-46.
65. Jadad AR, Moore RA, Carroll D, Jenkinson C, Reynolds DJ, Gavaghan DJ, et al. Assessing the quality of reports of randomized clinical trials: is blinding necessary? Control Clin Trials. 1996;17(1):1-12.
66. Probst P, Grummich K, Heger P, Zaschke S, Knebel P, Ulrich A, et al. Blinding in randomized controlled trials in general and abdominal surgery: protocol for a systematic review and empirical study. Syst Rev. 2016;5:48-.
67. Moher D, Hopewell S, Schulz KF, Montori V, Gøtzsche PC, Devereaux PJ, et al. CONSORT 2010 Explanation and Elaboration: updated guidelines for reporting parallel group randomised trials. BMJ (Clinical research ed). 2010;340:c869.
68. Moher D, Jones A, Lepage L, Group ftC. Use of the CONSORT Statement and Quality of Reports of Randomized TrialsA Comparative Before-and-After Evaluation. JAMA. 2001;285(15):1992-5.
69. Chalmers I, Oxman AD, Austvoll-Dahlgren A, Ryan-Vig S, Pannell S, Sewankambo N, et al. Key Concepts for Informed Health Choices: a framework for helping people learn how to assess treatment claims and make informed choices. BMJ evidence-based medicine. 2018;23(1):29-33.
70. Torgerson DJ, Roberts C. Understanding controlled trials. Randomisation methods: concealment. BMJ (Clinical research ed). 1999;319(7206):375-6.

71. Viera AJ, Bangdiwala SI. Eliminating bias in randomized controlled trials: importance of allocation concealment and masking. Fam Med. 2007;39(2):132-7.
72. Altman DG, Bland JM. Treatment allocation in controlled trials: why randomise? BMJ (Clinical research ed). 1999;318(7192):1209-.
73. Dettori J. The random allocation process: two things you need to know. Evidence-based spine-care journal. 2010;1(3):7-9.
74. Schulz KF, Grimes DA. Allocation concealment in randomised trials: defending against deciphering. The Lancet. 2002;359(9306):614-8.
75. Schulz KF, Chalmers I, Hayes RJ, Altman DG. Empirical Evidence of Bias: Dimensions of Methodological Quality Associated With Estimates of Treatment Effects in Controlled Trials. JAMA. 1995;273(5):408-12.
76. Savović J, Jones HE, Altman DG, Harris RJ, Jüni P, Pildal J, et al. Influence of reported study design characteristics on intervention effect estimates from randomized, controlled trials. Ann Intern Med. 2012;157(6):429-38.
77. Hewitt C, Hahn S, Torgerson DJ, Watson J, Bland JM. Adequacy and reporting of allocation concealment: review of recent trials published in four general medical journals. BMJ (Clinical research ed). 2005;330(7499):1057.
78. Clark L, Fairhurst C, Torgerson DJ. Allocation concealment in randomised controlled trials: are we getting better? BMJ (Clinical research ed). 2016;355:i5663.
79. Jüni P, Altman DG, Egger M. Assessing the quality of controlled clinical trials. BMJ : British Medical Journal. 2001;323(7303):42-6.
80. LeLorier J, Gregoire G, Benhaddad A, Lapierre J, Derderian F. Discrepancies between meta-analyses and subsequent large randomized, controlled trials. The New England journal of medicine. 1997;337(8):536-42.
81. Cappelleri JC, Ioannidis JP, Schmid CH, de Ferranti SD, Aubert M, Chalmers TC, et al. Large trials vs meta-analysis of smaller trials: how do their results compare? Jama. 1996;276(16):1332-8.
82. Hurwitz SR, Slawson DC. Should we be teaching information management instead of evidence-based medicine? Clin Orthop Relat Res. 2010;468(10):2633-9.
83. Boyack KW, Klavans R, Sorensen AA, Ioannidis JPA. A list of highly influential biomedical researchers, 1996–2011. Eur J Clin Invest. 2013;43(12):1339-65.
84. Slawson DC, Shaughnessy AF. Teaching evidence-based medicine: should we be teaching information management instead? Acad Med. 2005;80(7):685-9.
85. Sackett DL, Straus SE. Finding and applying evidence during clinical rounds: the "evidence cart". Jama. 1998;280(15):1336-8.
86. Haynes RB. Of studies, syntheses, synopses, summaries, and systems: the "5S" evolution of information services for evidence-based healthcare decisions. Evid Based Med. 2006;11(6):162-4.
87. Ebell MH, Sokol R, Lee A, Simons C, Early J. How good is the evidence to support primary care practice? Evidence Based Medicine. 2017;22(3):88-92.
88. Essential evidence plus Hoboken, N.J.: Wiley InterScience; 1996 [Available from: https://www.essentialevidenceplus.com/content/poems.

89. Grad R, Pluye P, Tang D, Shulha M, Slawson DC, Shaughnessy AF. Patient-oriented evidence that matters (POEMs) suggest potential clinical topics for the Choosing Wisely campaign. J Am Board Fam Med. 2015;28(2):184-9.
90. Group OLoEW. " The Oxford 2011 Levels of Evidence." Oxford Centre for Evidence-Based Medicine. http://wwwcebmnet/indexaspx?o=5653. 2011.
91. Shaughnessy AF, Slawson DC. Are we providing doctors with the training and tools for lifelong learning?. Interview by Abi Berger. BMJ (Clinical research ed). 1999;319(7220):1280-.
92. Shaughnessy AF, Slawson DC, Bennett JH. Becoming an information master: a guidebook to the medical information jungle. J Fam Pract. 1994;39(5):489-500.
93. Smith R. A POEM a week for the BMJ. BMJ (Clinical research ed). 2002;325(7371):983-.
94. Grad R, Practice Changers: Top 20 POEMs of 2018. FMX Family Medicine Experience; 2019.
95. Grad R, Pluye P, Johnson-Lafleur J, Granikov V, Shulha M, Bartlett G, et al. Do family physicians retrieve synopses of clinical research previously read as email alerts? J Med Internet Res. 2011;13(4):e101-e.
96. Focht III DR, Spicer C, Fairchok MP. The Efficacy of Duct Tape vs Cryotherapy in the Treatment of Verruca Vulgaris (the Common Wart). JAMA Pediatrics. 2002;156(10):971-4.
97. de Haen M, Spigt MG, van Uden CJT, van Neer P, Feron FJM, Knottnerus A. Efficacy of Duct Tape vs Placebo in the Treatment of Verruca Vulgaris (Warts) in Primary School Children. JAMA Pediatrics. 2006;160(11):1121-5.
98. Wenner R, Askari SK, Cham PMH, Kedrowski DA, Liu A, Warshaw EM. Duct Tape for the Treatment of Common Warts in Adults: A Double-blind Randomized Controlled Trial. JAMA Dermatology. 2007;143(3):309-13.
99. Boden WE, O'Rourke RA, Teo KK, Hartigan PM, Maron DJ, Kostuk WJ, et al. Optimal Medical Therapy with or without PCI for Stable Coronary Disease. New England Journal of Medicine. 2007;356(15):1503-16.
100. Prasad V, Ioannidis JP. Evidence-based de-implementation for contradicted, unproven, and aspiring healthcare practices. Implementation Science. 2014;9(1):1.
101. Mills EJ, Chan A-W, Wu P, Vail A, Guyatt GH, Altman DG. Design, analysis, and presentation of crossover trials. Trials. 2009;10:27-.
102. Krogh HB, Storebø OJ, Faltinsen E, Todorovac A, Ydedahl-Jensen E, Magnusson FL, et al. Methodological advantages and disadvantages of parallel and crossover randomised clinical trials on methylphenidate for attention deficit hyperactivity disorder: a systematic review and meta-analyses. BMJ Open. 2019;9(3):e026478.
103. Tice JA, Ettinger B, Ensrud K, Wallace R, Blackwell T, Cummings SR. Phytoestrogen supplements for the treatment of hot flashes: the Isoflavone Clover Extract (ICE) Study: a randomized controlled trial. Jama. 2003;290(2):207-14.
104. Krippendorff K. Content analysis: An introduction to its methodology: Sage; 2012.
105. Bateman ED, Boushey HA, Bousquet J, Busse WW, Clark TJH, Pauwels RA, et al. Can Guideline-defined Asthma Control Be Achieved? Am J Respir Crit Care Med. 2004;170(8):836-44.
106. Francis CW, Berkowitz SD, Comp PC, Lieberman JR, Ginsberg JS, Paiement G, et al. Comparison of ximelagatran with warfarin for the prevention of venous thromboembolism after total knee replacement. The New England journal of medicine. 2003;349(18):1703-12.

107. Gertsch JH, Basnyat B, Johnson EW, Onopa J, Holck PS. Randomised, double blind, placebo controlled comparison of ginkgo biloba and acetazolamide for prevention of acute mountain sickness among Himalayan trekkers: the prevention of high altitude illness trial (PHAIT). BMJ (Clinical research ed). 2004;328(7443):797.
108. Barrett-Connor E, Grady D, Sashegyi A, Anderson PW, Cox DA, Hoszowski K, et al. Raloxifene and cardiovascular events in osteoporotic postmenopausal women: four-year results from the MORE (Multiple Outcomes of Raloxifene Evaluation) randomized trial. Jama. 2002;287(7):847-57.
109. Wong WM, Wong BC, Hung WK, Yee YK, Yip AW, Szeto ML, et al. Double blind, randomised, placebo controlled study of four weeks of lansoprazole for the treatment of functional dyspepsia in Chinese patients. Gut. 2002;51(4):502-6.
110. Peter EA, Janssen PA, Grange CS, Douglas MJ. Ibuprofen versus acetaminophen with codeine for the relief of perineal pain after childbirth: a randomized controlled trial. CMAJ. 2001;165(9):1203-9.
111. Ely JW, Osheroff JA, Gorman PN, Ebell MH, Chambliss ML, Pifer EA, et al. A taxonomy of generic clinical questions: classification study. BMJ (Clinical research ed). 2000;321(7258):429-32.
112. Team RC. R: A language and environment for statistical computing, R Foundation for Statistical Computing Vienna, Austria. 2019 [Available from: https://www.R-project.org/.
113. Balachandran VP, Gonen M, Smith JJ, DeMatteo RP. Nomograms in oncology: more than meets the eye. The Lancet Oncology. 2015;16(4):e173-e80.
114. Gertler R, Stein HJ, Schuster T, Rondak IC, Hofler H, Feith M. Prevalence and topography of lymph node metastases in early esophageal and gastric cancer. Ann Surg. 2014;259(1):96-101.
115. Eyer F, Schuster T, Felgenhauer N, Pfab R, Strubel T, Saugel B, et al. Risk assessment of moderate to severe alcohol withdrawal--predictors for seizures and delirium tremens in the course of withdrawal. Alcohol Alcohol. 2011;46(4):427-33.
116. Frost J. The danger of overfitting regression models. Minitab Blog. 2015.
117. McNeish DM. Using Lasso for Predictor Selection and to Assuage Overfitting: A Method Long Overlooked in Behavioral Sciences. Multivariate Behavioral Research. 2015;50(5):471-84.
118. Breiman L. Statistical modeling: The two cultures (with comments and a rejoinder by the author). Statistical science. 2001;16(3):199-231.
119. Gerth WC, McCarroll KA, Santanello NC, Vandormael K, Zhang Q, Mannix LK. Patient satisfaction with rizatriptan versus other triptans: direct head-to-head comparisons. Int J Clin Pract. 2001;55(8):552-6.
120. Rosenfeldt V, Benfeldt E, Nielsen SD, Michaelsen KF, Jeppesen DL, Valerius NH, et al. Effect of probiotic Lactobacillus strains in children with atopic dermatitis. J Allergy Clin Immunol. 2003;111(2):389-95.
121. Nikander E, Kilkkinen A, Metsa-Heikkila M, Adlercreutz H, Pietinen P, Tiitinen A, et al. A randomized placebo-controlled crossover trial with phytoestrogens in treatment of menopause in breast cancer patients. Obstet Gynecol. 2003;101(6):1213-20.
122. Woessner KM, Simon RA, Stevenson DD. The safety of celecoxib in patients with aspirin-sensitive asthma. Arthritis Rheum. 2002;46(8):2201-6.

123. Sato Y, Honda Y, Iwamoto J, Kanoko T, Satoh K. Effect of folate and mecobalamin on hip fractures in patients with stroke: a randomized controlled trial. Jama. 2005;293(9):1082-8.
124. Bauchner H, Fontanarosa PB. Notice of Retraction: Sato Y, et al. Effect of Folate and Mecobalamin on Hip Fractures in Patients With Stroke: A Randomized Controlled Trial. JAMA. 2005;293(9):1082-1088.Notice of RetractionNotice of Retraction. JAMA. 2016;315(22):2405-.
125. Ioannidis JPA. Why Most Published Research Findings Are False. PLoS Med. 2005;2(8):e124.
126. Cook NR, Lee IM, Gaziano JM, Gordon D, Ridker PM, Manson JE, et al. Low-dose aspirin in the primary prevention of cancer: the Women's Health Study: a randomized controlled trial. Jama. 2005;294(1):47-55.
127. Ridker PM, Cook NR, Lee I-M, Gordon D, Gaziano JM, Manson JE, et al. A Randomized Trial of Low-Dose Aspirin in the Primary Prevention of Cardiovascular Disease in Women. New England Journal of Medicine. 2005;352(13):1293-304.
128. Miller FG, Joffe S. Equipoise and the Dilemma of Randomized Clinical Trials. New England Journal of Medicine. 2011;364(5):476-80.
129. Pocock SJ, Stone GW. The Primary Outcome Fails — What Next? New England Journal of Medicine. 2016;375(9):861-70.
130. Silberstein SD, Neto W, Schmitt J, Jacobs D, Group ftM-S. Topiramate in Migraine Prevention: Results of a Large Controlled Trial. JAMA Neurology. 2004;61(4):490-5.
131. Montori VM, Devereaux PJ, Adhikari NK, Burns KE, Eggert CH, Briel M, et al. Randomized trials stopped early for benefit: a systematic review. Jama. 2005;294(17):2203-9.
132. Bassler D, Briel M, Montori VM, Lane M, Glasziou P, Zhou Q, et al. Stopping randomized trials early for benefit and estimation of treatment effects: systematic review and meta-regression analysis. Jama. 2010;303(12):1180-7.
133. Bassler D, Montori VM, Briel M, Glasziou P, Walter SD, Ramsay T, et al. Reflections on meta-analyses involving trials stopped early for benefit: is there a problem and if so, what is it? Stat Methods Med Res. 2013;22(2):159-68.
134. Angiotensin-Converting–Enzyme Inhibition in Stable Coronary Artery Disease. New England Journal of Medicine. 2004;351(20):2058-68.
135. Prasad V. Plenary Session [a podcast on medicine, oncology, & health policy] [Internet]; 2018. Podcast. Available from: http://www.vinayakkprasad.com/plenarysession
136. Levels of Evidence Essential Evidence Plus [Available from: https://www-essentialevidenceplus-com.proxy3.library.mcgill.ca/product/ebm_loe.cfm?show=oxford.
137. Sibbald B, Roland M. Understanding controlled trials: Why are randomised controlled trials important? BMJ (Clinical research ed). 1998;316(7126):201.
138. Leichsenring F, Abbass A, Hilsenroth MJ, Leweke F, Luyten P, Keefe JR, et al. Biases in research: risk factors for non-replicability in psychotherapy and pharmacotherapy research. Psychol Med. 2017;47(6):1000-11.
139. Prasad V. Plenary Session [a podcast on medicine, oncology, & health policy] [Internet]; 2018. Podcast: 18:56. Available from: http://www.vinayakkprasad.com/plenarysession
140. Guyatt GH, Oxman AD, Vist GE, Kunz R, Falck-Ytter Y, Alonso-Coello P, et al. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. BMJ (Clinical research ed). 2008;336(7650):924-6.

Appendix

Appendix I – Example of POEM

Artificial hips and knees last up to 25 years

Published: 2019-04-04

Clinical question
How long do artificial hips and knees last?

Bottom line
Although the included studies do not take into account early failures, it appears that most hip and knee replacements last up to 25 years. (LOE = 2a)

Reference
Evans JT, Evans JP, Walker RW, Blom AW, Whitehouse MR, Sayers A. How long does a hip replacement last? A systematic review and meta-analysis of case series and national registry reports with more than 15 years of follow-up. Lancet 2019;393(10172):647-654. Evans JT, Walker RW, Evans JP, Blom AW, Sayers A, Whitehouse MR. How long does a knee replacement last? A systematic review and meta-analysis of case series and national registry reports with more than 15 years of follow-up. Lancet 2019;393(10172):655-663.

Study design: Meta-analysis (other)

Funding source: Government

Setting: Inpatient (any location) with outpatient follow-up

Synopsis
These researchers searched 2 databases and the bibliographies of included papers and review articles for pretty much any kind of study of hip or knee replacement surgery that reported at least 15 years of follow-up data. In addition, they also collected data from the national registries of Australia, Denmark, Finland, New Zealand, Norway, and Sweden. Those registries have all been around since 1975. The authors excluded papers not published in English. Although they did not clearly describe the process for selecting papers, the authors assessed the quality of included studies using a system adapted for the anticipated high rate of loss to follow-up. Ultimately, they included 44 hip replacement papers (13,212 hips) and 30 knee replacement papers (7232 knees). Additionally, the researchers had data from multiple national registry reports with more than 200,000 hip replacements and nearly 300,000 knee replacements. Among the studies of patients with knee replacements, none reported 25-year outcomes for patients with total knee replacements and only one reported 25-year outcomes for unicompartmental replacements. The average age of patients in the studies reporting hip replacement outcomes was younger than those in the national registries (58 years vs 68 years), but the average age of those undergoing knee replacement was comparable (~ 68 years). For hip replacements, the patients in the published studies fared better than those in the registries—78% and 58%, respectively, lasted 25 years. Among total knee replacements, the registry data indicate 82% survive 25 years. For unicompartmental replacements, the sole publication reported 72% last 25 years while the registry data showed 70% last that long. Because the authors excluded data with short-term follow-up, we have no information about early failures, and can only say that if the joint lasts 15 years, it is likely to last 25 years.

Henry C. Barry, MD, MS
Professor
Michigan State University
East Lansing, MI

Appendix II - Document 1 – Operational Definition and Coding Guide

POEM Rater Guide

Definition of MR

The concept of medical reversal (MR) occurs when new evidence finds an established medical practice to be less effective or more harmful than originally claimed, and contributes to practice change.

Operational definition

In the current study, a MR will be identified by comparing a POEM bottom line of a randomized controlled trial (RCT) with the latest evidence extracted from DynaMed.

Based on the above definition, an MR is when evidence shifts practice from green to red. However, we will also label POEMs where the evidence suggests a practice shift from red to green.

In addition, since evidence is fluid over time, each MR we identify will be further described in step 2 along a continuum. That being said, the final coding remains binary in step 1 (0- Not reversed; 1- Reversed).

Table 1 – MR Continuum (direction of the effect)

-2 (Strongly/really against) (red)	-1 (Weak against) negative waffle (orange)	+1 (Weak for) positive waffle (yellow)	+2 (Strongly/really for) (green)

A reversal would be:

- From (- or red) to (+ or green)
    - From -1 to +2
    - From -2 to +2
- From (+ or green) to (- or red)
    - From +1 to -2
    - From +2 to -2
- Situations where the evidence shifts from (+1) to (-1) or (-1) to (+1) also represent a reversal.

    - However, the 'certainty' of this reversal is 'weaker' than a situation where -1 turns into a 2. For this reason, there will be a second coding only for POEM-RCTs identified as reversed.
- Situations from -1 to -2 and from 1 to 2 do not constitute a reversal.

Table 2 – Suggestive Language Along the MR Continuum

-2 (Strongly/really against) (red)	-1 (Weak against) negative waffle (orange)	+1 (Weak for) positive waffle (yellow)	+2 (Strongly/really for) (green)
not effective OR ineffective	Likely ineffective	Likely effective/ beneficial	Significant reduction of [negative outcome]
Definitely do not use	Appears to be ineffective	Appears to be effective/ beneficial	Definitely should be used
Significant increase of [negative outcome]	Does not look like it can be effective	Looks like it can be effective	Significant increase of [positive outcome]
Significant decrease of [positive outcome]	Probably ineffective	Probably effective	Help
	May be ineffective	May be effective	Treats OR can treat

Second Coding

The second coding is applicable only when a POEM is labelled as contributing to an MR. The suggestive language described in table 2 serves as support to understand the MR continuum. This second coding will provide more information about the strength of the MR. Indeed, labeling a POEM as 'reversed' means there is new evidence showing an opposite direction of intervention effect (+ to - OR - to +). However, a simple label 'reversed/not reversed' is not enough to distinguish the different possible MRs across the continuum. For example, for a reversal from + to -, the second coding will help to distinguish whether the reversal was from +1 to -2, +1 to -1, +2 to -1, or +2 to -2.

Code Book for Data Extraction

Database Format

The database will be an Excel document with 8 columns. Each letter corresponds to a single column in our Excel document.

Here are the different columns with explanation:

A) **POEM number**
   - Number of the POEM in the list of 960 POEMs

B) **POEM ID Number**
   - Five-digit ID unique to each POEM

C) **POEM Title**
   - Full title of the POEM

D) **Full POEM (E1)**
   - The full POEM

E) **DynaMed (E2)**
   - Statement made by CR after looking at the updated evidence in DynaMed

F) **Question 1 - Is the bottom line of E1 reversed in 2018?**
   - Physician-raters' response

G) **Question 2 – Type of reversal**
   - Physician-raters' response to provide more information on POEMs that were identified as MR

H) **Comments/Notes by CR**
   - Comments and/or notes from CR to complement the information provided to the physician-raters

All columns are important for physician-raters except A and C.

Coding Sequence

1- Read all important information from a single line (*POEM title, Full POEM, DynaMed Statement, Comments/Notes by CR*);
2- With the information from *1-*, answer *Question 1* in the corresponding column *F*;
3- For MRs only, fill in the corresponding column *G* (*Second Coding*);
4- Repeat step 1- to 3- for all POEMs in the Excel document;
5- Once the process is complete, save and send back to CR.

Coding Guide (Operationalized in the Excel Spreadsheet)

Coding #1: Labeling POEMs as 'Reversed'

Question 1 (column F)

Is the bottom line of the POEM (E_i) reversed in 2018?

Possible answers

0: the bottom line of the POEM (E_i) is NOT reversed in 2018. When the updated evidence from DynaMed (or Essential Evidence Plus) is consistent with the bottom line of the POEM. When there is no new evidence contradicting the bottom line of the POEM for that condition in 2018.

1: the bottom line of the POEM (E_i) is reversed in 2018. When the updated evidence from DynaMed (or Essential Evidence Plus) is inconsistent with the bottom line of the POEM. When the updated evidence from DynaMed (or Essential Evidence Plus) suggests an opposite effect (+ to - OR - to +). See definition of MR for more detail. In the event that a POEM has multiple outcomes and we find one outcome to be reversed, then we consider the POEM to be a reversal. This POEM-RCT will be coded in step 2.

2: cannot be resolved; uncertain. For example, a POEM contains many outcomes with different or uncertain shifts of effect on practice.

3: when both 0 and 2 are true. For example, in the bottom line of one POEM, the effect of omalizumab is discussed. This POEM is 'not reversed' (0) as no subsequent and contradictory studies have been published, according to DynaMed. In addition, this POEM would also be seen as 'cannot be resolved' (2) as the drug was taken off the market. In such a case, please rate this POEM as a '3'.

Coding #2: For POEMs Labelled As Reversed

Question 2 (column G)

Since this POEM is an example of a medical reversal, how would you describe the type of reversal?

Possible answers

0: no reversal. When the bottom line of the POEM is still valid in 2018. This code was added to facilitate the use of this data in the different statistical approaches.

1: strong reversal (from +2 to -2). When the bottom line of the POEM is strongly/really for a practice and the updated evidence from DynaMed (or Essential Evidence Plus) is strongly/really against the same practice.

2: weak reversal (from +2 to -1). When the bottom line of the POEM is strongly/really for a practice and the updated evidence from DynaMed (or Essential Evidence Plus) is weakly/somewhat against the same practice.

3: weakest reversal (from +1 to -1). When the bottom line of the POEM is weakly/somewhat for a practice and the updated evidence from DynaMed (or Essential Evidence Plus) is weakly/somewhat against the same practice.

4: strong reversal (from -2 to +2). When the bottom line of the POEM is strongly/really against a practice and the updated evidence from DynaMed (or Essential Evidence Plus) is strongly/really for the same practice.

5: weak reversal (from -2 to +1). When the bottom line of the POEM is strongly/really against a practice and the updated evidence from DynaMed (or Essential Evidence Plus) is weakly/somewhat for the same practice.

6: weakest reversal (from -1 to +1). When the bottom line of the POEM is weakly/somewhat against a practice and the updated evidence from DynaMed (or Essential Evidence Plus) is weakly/somewhat for the same practice.

Appendix III – Coding Template Example

	POEM ID number	POEM Title	Full POEM	Dynamed statement	Is the bottom line of E1 reversed in 2018? 0 = no 1 = yes 2 = cannot be resolved; uncertain 3 = Both 0 and 2 apply	Type of reversal (see code book for details)	Comments/Notes by CR
261	60425	Eradicating HP reduces gastric CA risk	Does treatment of Helicobacter pylori infection reduce the risk of gastric cancer? Bottom line Asymptomatic carriers of Helicobacter pylori with no endoscopically determined precancerous gastric lesions are less likely to develop gastric cancer after eradication treatment. For most primary care clinicians, these patients will rarely, if ever, fall under their purview (most tests are ordered for symptomatic patients). We will need more evidence regarding long-term outcomes and cost/benefit analyses before we can justifiably screen all adults for H pylori infection. Reference Wong BC, Lam SK, Wong WM, et al. Helicobacter pylori eradication to prevent gastric cancer in a high-risk region of China. A randomized	The intervention was mentioned in DynaMed and no contradictory evidence was found. From Dynamed: "H. pylori eradication associated with reduced risk of gastric adenocarcinoma at 5 years after eradication, but not earlier (level 2 [mid-level] evidence)" AND "H. pylori eradication associated with reduced risk of gastric cancer in patients ≥ 60 years old (level 2 [mid-level] evidence)"	0	0	New evidence and consistent with the bottom line of the POEM
262	60427	Memantine + donepezil somewhat effective for AD	Is the combination of memantine and donepezil effective for the treatment of Alzheimer disease? Bottom line The combination of memantine (Namenda) and donepezil (Aricept) appears more beneficial in the treatment of Alzheimer disease than donepezil alone. Although the benefits were statistically significant when compared with placebo, it is difficult to assess the true clinical benefit of these minimal changes using various evaluation scoring tools. Evidence from trials evaluating more significant patient-oriented outcomes, such as delayed nursing home placement, is necessary before recommending widespread use of this treatment protocol. Reference Tariot PN, Farlow MR, Grossberg GT, Graham SM, McDonald S, Gergel I.	The intervention was mentioned in DynaMed and the following contradictory evidence was found « combination therapy with acetylcholinesterase inhibitors and memantine not associated with additional benefit compared to monotherapy for patients with mild to moderate Alzheimer dementia, and neither monotherapy nor combination therapy appear effective in patients with moderate to severe disease (level 2 [mid-level] evidence) based on systematic review with trial-specific quality measures not reported systematic review and network meta-analysis of 76 randomized trials evaluating memantine, acetylcholinesterase inhibitors, and combination therapy in 23,707 patients with Alzheimer dementia	1	weak reversal (from +1 to -2)	New evidence and inconsistent with the bottom line of the POEM.

Appendix IV – SuperType Table

SuperType	SuperType2	Comment
Ad	Ad	Practice administration / health systems
Dx	DxDf	Differential diagnosis
Dx	DxHP	Diagnosis - H&P (signs/symptoms)
Dx	DxTe	Diagnosis - lab tests
Dx	DxZA	Diagnosis - comparisons of above and miscellaneous
Ed	EdMD	Medical education
Ed	EdPt	Patient education
Et	EtCs	Causation and etiology
Et	EtEp	Incidence and prevalence of disease
Et	Etrk	Risk factors for disease
Px	Px	Prognosis of disease and natural history
Px	PxFU	Follow-up tests and monitoring
Sc	Sc	Screening for disease
Sc	ScPv	Prevention of disease (primary only)
Tx	TCAM	Treatment - complementary/alternative medicine
Tx	TxCt	Treatment - cost effectiveness and cost analyses
Tx	TxDt	Treatment - diet / exercise / PT / OT / vitamins
Tx	TxGd	Treatment - practice guidelines (can incl dx/tx)
Tx	TxHm	Treatment - harm (incl adverse drug effects)
Tx	TxRx	Treatment - drug therapy (incl surfactants, chemo)
Tx	TxSx	Treatment - surgery, anesthesia, andl procedures
Tx	TxZA	Treatment - comparisons of above, O2, behavioral, immunotx, other

Appendix V – R Commands

```
# read in data file
POEMdf<-read.csv('/Users/christianruchon/Desktop/POEM-DataAnalysis-
June2019II.csv',sep=';',header=T)
View(POEMdf)

# display variable names
names(POEMdf)

# descriptive stats for sample sizes
summary(POEMdf$Total.Sample.Size)

# descriptive stats for sample sizes
summary(POEMdf$Sample.Size.Intervention.Group)


# cases to  be removed because of 2 and 3 coding
dim(POEMdf)
which(POEMdf$`Q.1...Reversal.`==3|POEMdf$`Q.1...Reversal.`==2)
POEMdf<-POEMdf[-c(which(POEMdf$`Q.1...Reversal.`==3|POEMdf$`Q.1...Reversal.`==2)),]
dim(POEMdf)

POEMdf$Trial.Arms[which(POEMdf$Trial.Arms>3)]<-4
POEMdf$Trial.Arms<-as.factor(POEMdf$Trial.Arms)

# Confidence interval of binary outcomes for reversals of POEM-RCTs and POEM-Obs
binconf(34,408) #reversals of POEM-RCTs
binconf(12,51) #reversals of POEM-Obs

# frequency distribution of the outcome variable
table(POEMdf$Q.1...Reversal.)

#Generating a boxplot and histograms of trial sample size per outcome
par(mfrow=c(3,1))
boxplot(POEMdf$Total.Sample.Size~POEMdf$Q.1...Reversal., col=c("#00FFFF","#FF3399"),
        horizontal = TRUE, main = "POEM Total Sample Size Per Outcome",
        xlab="Trial Sample Size [axis log-scaled]", ylab = "", names=c("Not
Reversed","Reversed"), las=0, cex.lab=1.5,cex.main=1.5,log="x")

hist(POEMdf$Total.Sample.Size[which(POEMdf$Q.1...Reversal. == "0")],breaks =
seq(1,40000, by=100),xlab = "Trial Sample Size [axis log-scaled]",main = "POEM-RCTs
Identified as Not Reversed",
     col = "#00FFFF", cex.lab=1.5,cex.main=1.2,log="x")
legend("topright", inset=.02, legend=c("Not Reversed POEM-RCTs", "Reversed POEM-
RCTs"),col =c("#00FFFF", "#FF3399"),
       cex=1, fill=c("#00FFFF", "#FF3399"))
```

```
hist(POEMdf$Total.Sample.Size[which(POEMdf$Q.1...Reversal. == "1")],breaks =
seq(1,40000, by=150),xlab = "Trial Sample Size [axis log-scaled]",main = "POEM-RCTs
Identified as Reversed",
    col = "#FF3399", cex.lab=1.5, cex.main=1.2,log="x")


# fitting a conventional binary logistic regression to predict "medical reversal"
# outcome ¨MR yes/no¨
# explanatory variables: all remaining variables in the dataset
fitlogreg<-glm(`Q.1...Reversal.`~., family="binomial", data=POEMdf[,-c(2,3,9)])
summary(fitlogreg)


# as the conventional regression model is highly overfitted, we use so called lasso regression
instead
# this allows for entering many predictor variables in the model despite having only a relatively
low prevalence of the outcome

#https://stats.stackexchange.com/questions/72251/an-example-lasso-regression-using-glmnet-
for-binary-outcome

library(glmnet)

attach(POEMdf)
xfactors<-
model.matrix(Q.1...Reversal.~LOE+Trial.Arms+Allocation.Concealment+Drug.Non.Drug+Year
+agegroup+supertype+Setting)[,-1]
x<-as.matrix(data.frame(Total.Sample.Size, Sample.Size.Intervention.Group, xfactors))
fitglmnet<-glmnet(x, y=as.factor(Q.1...Reversal.), alpha=1, family="binomial")
#plot(fitglmnet, xvar="lambda",col=rainbow(16))

dev.off()
coef(fitglmnet)[, 10]
dotchart(coef(fitglmnet)[-1, 10])


# Nomogram
library(rms)
POEMdf$Year<-as.factor(POEMdf$Year)
Total.Sample.Size.Group<-
(Total.Sample.Size<100)*1+(Total.Sample.Size>=100&Total.Sample.Size<250)*2+(Total.Samp
le.Size>=250&Total.Sample.Size<500)*3+(Total.Sample.Size>=500)*4
boxplot(Total.Sample.Size~Total.Sample.Size.Group,log="y")
table(Total.Sample.Size.Group)
POEMdf<-cbind(POEMdf,Total.Sample.Size.Group)

ratio<-Sample.Size.Intervention.Group/Total.Sample.Size
POEMdf<-cbind(POEMdf,ratio)
```

```
table(POEMdf$LOE)
table(as.numeric(POEMdf$LOE))

LOE.cat<-
(as.numeric(POEMdf$LOE)==3|as.numeric(POEMdf$LOE)==4|as.numeric(POEMdf$LOE)==5
)*1
table(LOE,LOE.cat)

POEMdf<-cbind(POEMdf,LOE.cat)

ddist <- datadist(POEMdf[,-c(2,3)])
options(datadist='ddist')

f <-
lrm(Q.1...Reversal.~Total.Sample.Size.Group+ratio+LOE.cat+Trial.Arms+Allocation.Concealm
ent+Drug.Non.Drug+Year+agegroup+supertype+Setting,
       data=POEMdf)

summary (f)
exp(confint(f ))
exp(cbind(OR = coef(), confint(f)))
table(exp(cbind(OR = coef(f), confint(f))))
confint(f)
summary(model)$coefficients

f
summary (f)
exp(coef(f))
exp(confint.default(f))
table(summary (f))
nom <- nomogram(f, fun=function(x)1/(1+exp(-x)),  # or fun=plogis
            fun.at=c(.001,.01,.05,seq(.1,.9,by=.1),.95,.99,.999),
            funlabel="Estimated probability of medical reversal")

plot(nom)


f.final <- lrm(Q.1...Reversal.~LOE.cat+Drug.Non.Drug+Year,data=POEMdf)


nom.final <- nomogram(f.final, fun=function(x)1/(1+exp(-x)),  # or fun=plogis
            fun.at=c(.001,.01,.05,seq(.1,.9,by=.1),.95,.99,.999),
            funlabel="Estimated probability of medical reversal")

plot(nom.final)
```

```
xfactors<-
model.matrix(Q.1...Reversal.~LOE.cat+Total.Sample.Size.Group+Trial.Arms+Allocation.Conce
alment+Drug.Non.Drug+agegroup+supertype+Setting+Year)[,-1]
x<-as.matrix(data.frame(ratio, xfactors))
fitglmnet<-glmnet(x, y=as.factor(Q.1...Reversal.), alpha=1, family="binomial")
par(mfrow=c(1,1),mar=c(5,5,1,1))
#plot(fitglmnet, xvar="lambda",col=rainbow(12),lwd=c(1,1,3,1,1,1,1,3,1,1,1,5))
#legend("bottomright",legend=names(coef(fitglmnet)[, 10]),
#       lty=rep(1,12),col=rainbow(12),cex=0.5)
#intercept<-0
#names(intercept)<-"intercept"
dotchart(coef(fitglmnet)[-1, 10],pch=20,cex=1.5,xlab="shrinked regression coefficients [lasso
regression]")
abline(v=0,lty=2)

coef(fitglmnet)[, 10]


# Random Forest

table(Q.1...Reversal.)

extend<-numeric(0)
for(i in 1:10)
{
extend<-rbind(extend,POEMdf[POEMdf$Q.1...Reversal.==1,])
}

dim(extend)

POEMdf.W<-rbind(POEMdf,extend)
dim(POEMdf.W)
table(POEMdf.W$Q.1...Reversal.)

POEMdf.W$agegroup<-as.factor(POEMdf.W$agegroup)
POEMdf.W$LOE.cat<-as.factor(POEMdf.W$LOE.cat)
POEMdf.W$Setting<-as.factor(POEMdf.W$Setting)
POEMdf.W$Total.Sample.Size.Group<-as.factor(POEMdf.W$Total.Sample.Size.Group)
POEMdf.W$Trial.Arms<-as.factor(POEMdf.W$Trial.Arms)
POEMdf.W$supertype<-as.factor(POEMdf.W$supertype)
POEMdf.W$Drug.Non.Drug<-as.factor(POEMdf.W$Drug.Non.Drug)

library(randomForest)

set.seed(1234)
randvar<-rnorm(nrow(POEMdf.W)) # Random normally distributed variable generated
randvar2<-rbinom(nrow(POEMdf.W),1,0.5) # Random binary variable generated
POEMdf.W<-cbind(POEMdf.W,randvar,randvar2)
```

```
rf<-randomForest(factor(Q.1...Reversal.)~.,data=POEMdf.W[,-
c(2,3,9,4,5,6)],ntree=1000,importance=T)
rf
varImpPlot(rf)

# classification tree
library(rpart)
library(rpart.plot)

tree<-rpart(factor(Q.1...Reversal.)~.,data=POEMdf.W[,-c(2,3,9,4,5,6)],cp=0.02,maxdepth =3)
rpart.plot(tree,box.palette = "RdBu",shadow.col = "gray",nn=T, main="Classification tree")


# Random Forest with the 3 variables confirmed with the lasso regression
rf.final<-
randomForest(factor(Q.1...Reversal.)~LOE.cat+Drug.Non.Drug+Year,data=POEMdf.W[,-
c(2,3,9,4,5,6)],ntree=1000,importance=T)
rf.final
varImpPlot(rf.final)

# classification tree with the 3 variables confirmed with the lasso regression
tree.final<-rpart(factor(Q.1...Reversal.)~LOE.cat+Drug.Non.Drug+Year,data=POEMdf.W[,-
c(2,3,9,4,5,6)],cp=0.02,maxdepth =3)
rpart.plot(tree.final,box.palette = "RdBu",shadow.col = "gray",nn=T, main="Classification tree
of the 3 variables of the lasso regression")
```

Appendix VI –POEM-RCTs identified as reversed

	POEM-RCT (E1) POEM title; PMID	E2 Source; PMID	Direction	Comment
1	Sulindac does not prevent polyps in familial adenomatous polyposis; 11932472	Kim, B., & Giardiello, F. M. (2011). Chemoprevention in familial adenomatous polyposis. Best practice & research Clinical gastroenterology, 25(4-5), 607-622. ; 22122775	Weakest reversal (from -1 to +1)	Nonsteroidal anti-inflammatory drugs (NSAIDs) (celecoxib or sulindac) reduce number of polyps in patients with familial adenomatous polyposis (FAP), but effect on cancer risk unknown (level 3 [lacking direct] evidence). sulindac for primary prevention (in patients with abnormal gene) not associated with significant difference in adenomas in 1 trial. in 3 trials of secondary prevention, sulindac and celecoxib each reduced relative number of colorectal adenomas (range of reduction of polyp burden of 11.9%-44% with NSAID vs. 4.5%-10% with control).
2	Magnesium effective in severe, acute asthma; 12171821	Kew, K. M., Kirtchuk, L., & Michell, C. I. (2014). Intravenous magnesium sulfate for treating adults with acute asthma in the emergency department. Cochrane Database of Systematic Reviews, (5). ; 24865567	Weakest reversal (from -1 to +1)	Raters provided no additional comment.

3	HSV vaccine is safe and effective; 12444179	Hensel, M. T., Marshall, J. D., Dorwart, M. R., Heeke, D. S., Rao, E., Tummala, P., ... & Sloan, D. D. (2017). Prophylactic herpes simplex virus 2 (HSV-2) vaccines Adjuvanted with stable emulsion and toll-like receptor 9 agonist induce a robust HSV-2-specific cell-mediated immune response, protect against symptomatic disease, and reduce the latent viral reservoir. Journal of virology, 91(9), e02257-16. ; 28228587	Weak reversal (from +2 to -1)	The primary endpoint of disease prevention in seropositive individuals was not achieved.
4	Oral levonorgestrel = mifepristone for emergency contraception; 12480356	Shen, J., Che, Y., Showell, E., Chen, K., & Cheng, L. (2017). Interventions for emergency contraception. Cochrane Database of Systematic Reviews, (8). ; 28766313	Weak reversal (from +2 to -1)	Raters provided no additional comment.
5	Nebulized 3% saline more effective for viral bronchiolitis; 12475841	Silver, A. H., Esteban-Cruciani, N., Azzarone, G., Douglas, L. C., Lee, D. S., Liewehr, S., ... & Rinke, M. L. (2015). 3% Hypertonic saline versus normal saline in inpatient bronchiolitis: a randomized controlled trial. Pediatrics, 136(6), 1036-1043. ; 26553190	Weakest reversal (from +1 to -1)	Raters provided no additional comment.

6	Fetal ECG reduces neonatal encephalopathy; 12548215	Brocklehurst, P., Field, D., Greene, K., Juszczak, E., Kenyon, S., Linsell, L., ... & Steer, P. (2018). Computerised interpretation of the fetal heart rate during labour: a randomised controlled trial (INFANT). Health Technology Assessment, 22(9). ; 29437032	Weakest reversal (from +1 to -1)	Conclusion of the trial (E2) : "This trial does not support the hypothesis that the use of computerised interpretation of the CTG in women who have EFM in labour improves the clinical outcomes for mothers or babies".
7	Steroids ineffective for throat pain in children; 12712025	Hayward, G., Thompson, M., Heneghan, C., Perera, R., Del Mar, C., & Glasziou, P. (2009). Corticosteroids for pain relief in sore throat: systematic review and meta-analysis. Bmj, 339, b2976. ; 19661138	Weakest reversal (from -1 to +1)	The cited evidence suggest that corticosteroids provide symptomatic relief of pain in sore throat.
8	Isoflavones not very effective for hot flashes; 12738504	Taku, K., Melby, M. K., Kronenberg, F., Kurzer, M. S., & Messina, M. (2012). Extracted or synthesized soybean isoflavones reduce menopausal hot flash frequency and severity: systematic review and meta-analysis of randomized controlled trials. Menopause, 19(7), 776-790. ; 22433977	Weakest reversal (from -1 to +1)	From the SR published in Menopause: "mean placebo-subtracted reduction with isoflavones ranged from 3% to 57% results were statistically significant in 10 of 13 trials".

9	Treating BV with intravaginal clinda reduces preterm births; 12636956	Subtil, D., Brabant, G., Tilloy, E., Devos, P., Canis, F., Fruchart, A., ... & Gautier, S. (2018). Early clindamycin for bacterial vaginosis in pregnancy (PREMEVA): a multicentre, double-blind, randomised controlled trial. The Lancet, 392(10160), 2171-2179. ; 30322724	Weak reversal (from +2 to -1)	The new PREMEVA trial recommends weakly/somewhat against the practice of clindamycin for BV in pregnancy.
10	Tubes for otitis ineffective for language development; 12897272	Steele, D. W., Adam, G. P., Di, M., Halladay, C. H., Balk, E. M., & Trikalinos, T. A. (2017). Effectiveness of tympanostomy tubes for otitis media: a meta-analysis. Pediatrics, 139(6), e20170125. ; 28562283	Weak reversal (from -2 to +1)	This is subject to change as new evidence emerges for benefits at 12 to 24 months
11	Shock wave therapy ineffective for plantar fasciitis; 12855524	Li, H., Xiong, Y., Zhou, W., Liu, Y., Liu, J., Xue, H., ... & Liu, G. (2019). Shock-wave therapy improved outcome with plantar fasciitis: a meta-analysis of randomized controlled trials. Archives of orthopaedic and trauma surgery, 1-8. ; 31435724	Weak reversal (from -2 to +1)	Raters provided no additional comment.
12	Duloxetine safe, effective for stress urinary incontinence; 14501737	Maund, E., Guski, L. S., & Gøtzsche, P. C. (2017). Considering benefits and harms of duloxetine for treatment of stress urinary incontinence: a meta-analysis of clinical study reports. Cmaj, 189(5), E194-E203. ; 28246265	Weak reversal (from +2 to -1)	From the cited meta-analysis: "although duloxetine is effective for stress urinary incontinence in women, the rates of associated harm were high when individual patient data were analyzed, and the harms outweighed the benefits."

13	Flu vaccine not effective for preventing AOM in kids < 24 months, 14506120	Norhayati, M. N., Ho, J. J., & Azman, M. Y. (2017). Influenza vaccines for preventing acute otitis media in infants and children. Cochrane Database of Systematic Reviews, (10). ; 29039160	Weak reversal (from -2 to +1)	Raters provided no additional comment.
14	Fluconazole improves septic shock outcomes; 12847386	Schuster, M. G., Edwards, J. E., Sobel, J. D., Darouiche, R. O., Karchmer, A. W., Hadley, S., ... & Rex, J. H. (2008). Empirical fluconazole versus placebo for intensive care unit patients: a randomized trial. Annals of internal medicine, 149(2), 83-90. ; 18626047	Weakest reversal (from +1 to -1)	Raters provided no additional comment.
15	Iron supplement in pregnancy may improve birthweight; 14522736	Ziaei, S., Norrozi, M., Faghihzadeh, S., & Jafarbegloo, E. (2007). A randomised placebo-controlled trial to determine the effect of iron supplementation on pregnancy outcome in pregnant women with haemoglobin≥ 13.2 g/dl. BJOG: An International Journal of Obstetrics & Gynaecology, 114(6), 684-688. ; 17516958	Strong reversal (from -2 to +2)	Raters provided no additional comment.
16	Sibutramine effective in the treatment of binge-eating disorder; 14609886	Tziomalos, K., Krassas, G. E., & Tzotzas, T. (2009). The use of sibutramine in the management of obesity and related disorders: an update. Vascular health and risk management, 5, 441. ; 19475780	Weakest reversal (from +1 to -1)	Sibutramine remains effective, however it is associated with side effects and its use is restricted. Here, the reversal results from the drug being more harmful than originally claimed.

17	Efalizumab effective in plaque psoriasis; 14627785	Lin, E. J., Reddy, S., Shah, V. V., & Wu, J. J. (2018). A review of neurologic complications of biologic therapy in plaque psoriasis. Cutis, 101(1), 57-60. ; 29529105	Weak reversal (from +2 to -1)	Efalizumab is more harmful than originally thought and was withdrawn from the market in 2009 for causing progressive multifocal leukoencephalopathy (PML).
18	Vasopressin better than epi for asystole; 14711909	Gueugniaud, P. Y., David, J. S., Chanzy, E., Hubert, H., Dubien, P. Y., Mauriaucourt, P., ... & Thiercelin, D. (2008). Vasopressin and epinephrine vs. epinephrine alone in cardiopulmonary resuscitation. New England journal of medicine, 359(1), 21-30. ; 18596271	Strong reversal (from +2 to -2)	Vasopressin seems to offer no advantage either as substitute for or in combination with standard-dose epinephrine in cardiac arrest.
19	Memantine + donepezil somewhat effective for AD; 14734594	Tsoi, K. K., Chan, J. Y., Chan, F. C., Hirai, H. W., Kwok, T. C., & Wong, S. Y. (2019). Monotherapy Is Good Enough for Patients with Mild-to-Moderate Alzheimer's Disease: A Network Meta-analysis of 76 Randomized Controlled Trials. Clinical Pharmacology & Therapeutics, 105(1), 121-130. ; 29717478	Weakest reversal (from +1 to -1)	Raters provided no additional comment.

20	Lidocaine + naproxen reduces pain w/ endometrial sampling; 14754707	Api, O., Ergen, B., Api, M., Ugurel, V., Emeksiz, M. B., & Unal, O. (2010). Comparison of oral nonsteroidal analgesic and intrauterine local anesthetic for pain relief in uterine fractional curettage: a randomized, double-blind, placebo-controlled trial. American journal of obstetrics and gynecology, 203(1), 28-e1. ; 20435293	Weakest reversal (from +1 to -1)	Small, likely underpowered study showed lido + another NSAID no more effective than either alone.
21	HP eradication may cause minor worsening of GERD; 14724146	Saad, A. M., Choudhary, A., & Bechtold, M. L. (2012). Effect of Helicobacter pylori treatment on gastroesophageal reflux disease (GERD): meta-analysis of randomized controlled trials. Scandinavian journal of gastroenterology, 47(2), 129-135. ; 22229305	Weakest reversal (from +1 to -1)	Raters provided no additional comment.
22	Valproate may be effective for painful diabetic neuropathy; 14702509	Griebeler, M. L., Morey-Vargas, O. L., Brito, J. P., Tsapas, A., Wang, Z., Leon, B. G. C., ... & Murad, M. H. (2014). Pharmacologic interventions for painful diabetic neuropathy: an umbrella systematic review and comparative effectiveness network meta-analysis. Annals of Internal Medicine, 161(9), 639-649. 25364885	Weakest reversal (from -1 to +1)	Raters provided no additional comment.

23	Seizure prophylaxis unnecessary in childhood acute head injury; 15039684	Thompson, K., Pohlmann-Eden, B., Campbell, L. A., & Abel, H. (2015). Pharmacological treatments for preventing epilepsy following traumatic head injury. Cochrane database of systematic reviews, (8). ; 26259048	Weakest reversal (from -1 to +1)	Raters provided no additional comment.
24	Soy protein isoflavones do not reduce menopausal complications; 15238592	Cheng, P. F., Chen, J. J., Zhou, X. Y., Ren, Y. F., Huang, W., Zhou, J. J., & Xie, P. (2015). Do soy isoflavones improve cognitive function in postmenopausal women? A meta-analysis. Menopause, 22(2), 198-206. ; 25003621	Weakest reversal (from -1 to +1)	New evidence shows benefit on cognitive function.
25	SSRIs ineffective for hot flashes; 15668596	Handley, A. P., & Williams, M. (2015). The efficacy and tolerability of SSRI/SNRIs in the treatment of vasomotor symptoms in menopausal women: a systematic review. Journal of the American Association of Nurse Practitioners, 27(1), 54-61. ; 24944075	Weak reversal (from -2 to +1)	Raters provided no additional comment.
26	Recombinant factor VIIa improves ICH outcomes; 15728810	Salman, R. A. S., Law, Z. K., Bath, P. M., Steiner, T., & Sprigg, N. (2018). Haemostatic therapies for acute spontaneous intracerebral haemorrhage. Cochrane Database of Systematic Reviews, (4). ; 29664991	Weak reversal (from +2 to -1)	Raters provided no additional comment.

27	Tegaserod effective for chronic constipation; 15667494	Evans, B. W., Clark, W. K., Moore, D. J., & Whorwell, P. J. (2007). Tegaserod for the treatment of irritable bowel syndrome and chronic constipation. Cochrane database of systematic reviews, (4). ; 17943807	Weak reversal (from +2 to -1)	The evidence in DynaMed:"tegaserod improves symptoms for women with IBS with mixed bowel habits or constipation (level 1 [likely reliable] evidence)" suggests no contraindication. "tegaserod is effective for women with constipation-predominant IBS initially and during recurrence (level 1 [likely reliable] evidence)." It is no longer considered a safe treatment due to cardiovascular effects, hence it was discontinued.
28	Functional magnetic stimulation reduces mixed urinary incontinence; 15821527	Suzuki, T., Yasuda, K., Yamanishi, T., Kitahara, S., Nakai, H., Suda, S., & Ohkawa, H. (2007). Randomized, double-blind, sham-controlled evaluation of the effect of functional continuous magnetic stimulation in patients with urgency incontinence. Neurourology and Urodynamics: Official Journal of the International Continence Society, 26(6), 767-772. ; 17397061	Strong reversal (from +2 to -2)	Double blind RCT indicating this intervention is no more effective than sham therapy.
29	Nicotine patch probably effective in adolescents; 14606993	Fanshawe, T. R., Halliwell, W., Lindson, N., Aveyard, P., Livingstone-Banks, J., & Hartmann-Boyce, J. (2017). Tobacco cessation interventions for young people. Cochrane Database of Systematic Reviews, (11). ; 29148565	Weakest reversal (from +1 to -1)	Raters provided no additional comment.

30	N-acetylcysteine ineffective in COPD (BRONCUS); 15866309	Poole, P., Sathananthan, K., & Fortescue, R. (2019). Mucolytic agents versus placebo for chronic bronchitis or chronic obstructive pulmonary disease. Cochrane Database of Systematic Reviews, (5). ; 31107966	Weak reversal (from -2 to +1)	Given the DynaMed article gives level 2 evidence and states mucolytics "might reduce exacerbations, hospitalizations…" this would suggest weakly positive evidence to support the intervention.
31	Low-dose aspirin doesn't lower women's cancer risk (WHS); 15998890	Cook, N. R., Lee, I. M., Zhang, S. M., Moorthy, M. V., & Buring, J. E. (2013). Alternate-day low-dose aspirin and cancer risk: long-term observational follow-up of a randomized trial. Annals of internal medicine, 159(2), 77. ; 23856681	Weakest reversal (from -1 to +1)	The reversal is specific to the reduced risk of colorectal cancer in women. As per the study sited: « Long-term use of alternate-day, low-dose aspirin may reduce risk for colorectal cancer in healthy women ».
32	Acupuncture ineffective for fibromyalgia; 15998750	Uğurlu, F. G., Sezer, N., Aktekin, L., Fidan, F., Tok, F., & Akkuş, S. (2017). The effects of acupuncture versus sham acupuncture in the treatment of fibromyalgia: a randomized controlled clinical trial. Acta reumatologica portuguesa, (1). ; 28371571	Weak reversal (from -2 to +1)	Acupuncture significantly improved pain and symptoms of fibromyalgia. Although sham effect was important, real acupuncture treatment seems to be effective in treatment of fibromyalgia.

33	Levofloxacin reduces hospitalization but not death in chemo patients; 16148284	Gafter-Gvili, A., Fraser, A., Paul, M., Vidal, L., Lawrie, T. A., van de Wetering, M. D., ... & Leibovici, L. (2012). Antibiotic prophylaxis for bacterial infections in afebrile neutropenic patients following chemotherapy. Cochrane database of systematic reviews, (1). ; 22258955	Strong reversal (from -2 to +2)	A Cochrane 2012 study on febrile neutropenia, also cited on DynaMed under 'Febrile Neutropenia', showed level 1 evidence in favor of antibiotic prophylaxis for pts with afebrile neutropenia after chemo.
34	Newer antipsychotics similar to older agents (CATIE); 16172203	De Berardis, D., Rapini, G., Olivieri, L., Di Nicola, D., Tomasetti, C., Valchera, A., ... & Serafini, G. (2018). Safety of antipsychotics for the treatment of schizophrenia: a focus on the adverse effects of clozapine. Therapeutic advances in drug safety, 9(5), 237-256. ; 29796248	Weakest reversal (from -1 to +1)	Per DynaMed on Medications for schizophrenia, "clozapine associated with reduced risk for suicide death vs. perphenazine (HR 0.34, 95% CI 0.2-0.57)", and "antipsychotic use, particularly clozapine, associated with decreased mortality in patients with schizophrenia (level 2 [mid-level] evidence)." This suggests that a newer agent has benefits over an older agent, which would be a weaker reversal of the POEM suggesting that older agents are somewhat preferred.