Meta-epidemiologic consideration of confounding for health care decision making

Inauguraldissertation

zur

Erlangung der Würde eines Doktors der Philosophie

vorgelegt der

Philosophisch-Naturwissenschaftlichen Fakultät

der Universität Basel

von

Hannah Ewald

aus Amberg, Deutschland

Basel, 2018

Genehmigt von der Philosophisch-Naturwissenschaftlichen Fakultät

auf Antrag von

Prof. Dr. Marcel Tanner (Fakultätsverantwortlicher)

PD Dr. med. Lars G. Hemkens (Dissertationsleiter)

Prof. Dr. med. Atle Fretheim (Koreferent)

Basel, den 27. März 2018

Prof. Dr. Martin Spiess (Dekan)

Table of contents

Acknowledgements

I would like to thank everyone who helped realize this project.

Special thanks go to:

Lars G Hemkens, for excellent supervision and great humor, who never failed to notice my nightly rides around the office on a magical broomstick or my sophisticated conversations with the office's unofficial pet raven.

Heiner C Bucher, for giving me the opportunity to do my PhD at the Institute for Clinical Epidemiology & Biostatistics, and to "pack my suitcase full of scientific tools".

John PA Ioannidis, for sharing his knowledge and insight and making me feel part of a much bigger picture.

Aviv Ladanie, my smart co-PhD student, whose help in coding saved me from premature-jumping-out-of-the-window.

Dominik Glinz, for reminding me that coffee breaks have a purpose – even if I don't drink coffee.

Kimberly Mc Cord, for secretly providing me with the best chocolate chip cookies in the world when I'm not trying to enhance my brain function with a zero-added-sugar diet.

Mirco Wedel, for his unconditional support and inspiring discussions on statistical issues, even at 4 in the morning.

My family and friends, for believing in me.

The world is a better place with you all in it.

Abbreviations

ACTG	AIDS Clinical Trial Group
AIDS	Acquired Immune Deficiency Syndrome
ARISTOTLE	Apixaban for Reduction in Stroke and Other Thromboembolic Events in Atrial Fibrillation trial
ART	Anti-retroviral therapy
ATHENA	AIDS Therapy Evaluation Netherlands
BL	Baseline
BMJ	The British Journal of Medicine
CALERIE	Comprehensive Assessment of the Long-term Effects of Reducing Intake of Energy
CD4	Cluster of differentiation 4
CDC-C	Centers for Disease Control and Prevention classification system for HIV-infection, category C: severely symptomatic
CEB	Basel Institute for Clinical Epidemiology and Biostatistics
CER	Comparative effectiveness research
CHD	Coronary heart disease
CI	Confidence interval
CMAJ	Canadian Medical Association Journal
COREYA study	COhort with REYAtaz study
CoRIS	Cohorte de la Red de Investigación en SIDA
CVD	Cardiovascular disease
DOPPS	Dialysis Outcomes and Practice Patterns Study
EQUATOR	Enhancing the QUAlity and Transparency Of health Research
ESRD	End stage renal disease
EVOO	Extra virgin olive oil
FHDH-ANRS CO4	French Hospital Database on HIV—Agence Nationale de Recherches sur le SIDA
GEMES	Grupo Español Multicéntrico para el Estudio de Seroconvertores-Haemophilia
GPRD	General Practice Research Database
GRADE	Grading of Recommendations Assessment, Development and Evaluation
HAART	Highly active anti-retroviral therapy
HIV	Human immunodeficiency virus
HR	Hazard ratio
ICH	International Conference on Harmonisation
IF	Impact factor
IPCW	Inverse probability of censoring weighting
IPTW	Inverse probability of treatment weighting
IPW	Inverse probability weighting
IQR	Interquartile range
ISAARV	Initiative Sénégalaise d'Accès aux Medicaments Anti-rétroviraux
ITT	Intention-to-treat
JAMA	Journal of the American Medical Association

MACS	Multicenter AIDS Cohort Study
MEDLINE	Medical Literature Analysis and Retrieval System Online
MSM	Marginal structural models
NA	Not applicable
NSCLC	Non-small-cell lung cancer
OEDTR	Austrian Dialysis and Transplant Registry
OR	Odds ratio
PHS	Physicians' Health Study
PICO	Patient – Intervention – Comparison – Outcome
PISCIS	Proyecto para la Informatización del Seguimiento Clinico-epidemiológico de la Infección por HIV y SIDA
PLOS	Public Library of Science
PP	Per protocol
PPHS	PhD Educational Platform Health Sciences
PREDIMED	Primary Prevention of Cardiovascular Disease with a Mediterranean Diet
RCT	Randomized controlled trial
RECORD	REporting of studies Conducted using Observational Routinely collected Data
ROBINS	Risk Of Bias In Non-randomised Studies of Interventions
ROR	Ratio of odds ratios
SHCS	Swiss HIV Cohort Study
SNMs	Structural nested models
STROBE	Strengthening the Reporting of Observational Studies in Epidemiology
Swiss TPH	Swiss Tropical and Public Health Institute
THIN	The Health Improvement Network
UK	United Kingdom
UK CHIC	United Kingdom Collaborative HIV Cohort
US VACS-VC	United States Veterans Aging Cohort Study–Virtual Cohort
USA	United States of America
USRDS	US Renal Data System
WHI	Women's Health Initiative
WHS	Women's Health Study

Plain language summary

As patients, we all want to believe that there is the right medical solution for every ailment and that our doctor knows best. What we usually don't know is that our doctor's knowledge is based on experience and on evidence. However, the evidence can be flawed, exaggerated, or may not actually apply to us. While there are many things that can go wrong in clinical studies, the main focus of this dissertation is on the concept of confounding. Confounding occurs when a specific exposure and outcome have a common cause. For example, more breast cancer patients receiving surgery as the observed "exposure" survive than those receiving chemotherapy. Concluding that surgery is better for survival may, however, be confounded by cancer stage because those who were operated on had a less advanced cancer stage and thus were more likely to survive to begin with. Minimizing the impact of such confounding in research studies on treatment effects is important because it can alter the estimates of a treatment effect and thus may lead to wrong conclusions and ultimately to wrong treatment decisions.

For many health topics, there are myriads of studies available and whether or not their results can give us reliable answers to what we want to know depends on a variety of factors. The most important factor is the study design. Randomized controlled trials (RCTs) are the current gold standard to produce evidence for treatment decisions. They measure the causal effect of a treatment versus a control on a specific outcome. The key element is that study participants are randomly assigned to treatment or control (which could be a placebo or another treatment). The randomization tries to balance all known (such as age) and unknown characteristics (such as undiagnosed diseases) of the participants which means that they also balance all known and unknown confounding factors. The only difference between the participant groups will then be the allocation to a treatment or control. This would be the perfect study design if the circumstances were ideal, i.e. if every participant adhered to the assigned treatment and stayed on the study until the end. In reality, the participants often do not adhere (e.g. because the exercise program of a weight-loss study is too demanding) or they become lost to follow-up (e.g. because they moved away or did not want to be on the study anymore). However, not every clinical question can be answered in an RCT. Another important research design are observational studies, where the exposure of patients to an intervention or a control is not decided by the study investigators (thus observational) and may thus depend on a number of other known and unknown factors, e.g. doctors' decisions or patient's preferences. This study design is very prone to confounding and requires careful statistical analyses. Statistical methods can then be used to retrospectively address issues like confounding or confounding that changes over time. One such statistical method is marginal structural models (MSM). MSM allow a causal interpretation of results under the assumptions that all confounding factors are known, correctly measured and properly implemented in the statistical models. However, even with the latest statistical methods, RCTs and observational studies may not give the same answer when trying to solve the same question. Hence, the aims of the doctoral projects were 1) to evaluate the extent to which confounding is actively considered in the conclusions from observational studies; 2) to evaluate the agreement of treatment effects from non-randomized studies using MSM with reported effects from RCTs on the same topic; 3) to evaluate when MSM is used in RCTs and how these results differ from the main (non-MSM) results of the same trial.

First, we assessed the scope of the issue within the health professionals' literature. Are authors of scientific papers aware of the problem of confounding for the interpretation of their results and do they present their results in light of its possible impact? Second, if observational studies use MSM to

reduce the impact of confounding and allow a causal interpretation, the results should be similar to those from RCTs on the same clinical question. To assess how well they agree, we used established approaches to compare the effects, for example we determined how often the effects from both designs indicated concordantly that a treatment is beneficial or not. Third, we conducted an empirical analysis of where and why MSM is used to analyze randomized comparisons, a rather new and emerging approach to address confounding within randomized trials, and how these results compare to non-MSM results from the same trial.

We found that observational studies in general tend to have unsatisfactory or no discussion of confounding at all. If confounding was mentioned, it was either deemed irrelevant for the respective research or results are not brought in context of necessary cautious interpretation. Studies that did, however, report possible limitations due to confounding were actually cited more by other researchers than studies that deemed an influence due to confounding unlikely. This means research that is carefully reported may have more impact on science than other research.

When MSM was applied to observational study data, the effects often had opposite directions (i.e. one showed harm and the other benefit of the intervention) and were more favorable for the experimental treatment than in randomized studies on the same research question. This was even more so when the studies focused on informing health care decision making rather than statistical methodology.

MSM was applied to RCTs to minimize the influence of confounding that arises when study participants do not adhere to the protocol. Within the main publication and the publication reporting MSM-based results (sometimes the same), authors reported on average 6 analyses for one outcome in the same population and at the same point in time. Most of these results, however, pointed in the same direction and had more or less similar effect sizes, which means that the clinical interpretation is often similar.

We can never be certain that we know all confounding factors, measured them correctly and implemented them correctly in the statistical models. Even research that used causal modelling techniques may still come to different answers than RCTs evaluating the same clinical question would. Hence, confounding should be more carefully acknowledged in non-randomized research, doing so is not associated with lower citation impact. Results from causal modelling can be useful sensitivity analyses that can help researchers to get a bigger picture of the impact of other influencing factors. Health care decision makers should remain cautious when using non-randomized evidence to guide their health care decisions.

Introduction

In evidence-based medicine, the best available clinical evidence, the clinician's expertise, and the patient's values and preferences are applied to make an individualized, medical, evidence-based decision[1]. Study designs to gather evidence are frequently classified into two main categories: randomized controlled trials (RCTs), where patients are randomly allocated to an intervention or a control, and observational studies, where the exposure of patients to an intervention or a control is not controlled by the investigators (thus observational) and may thus depend on a number of other known and unknown factors, e.g. doctors' decisions or patients' preferences[2]. The estimated effects from both study designs can deviate from the true (and unknown) effect of the treatments for many reasons including random error and bias. The key advantage of RCTs over observational studies is better control of a number of biases.

Bias is any error that leads to the systematic over- or underestimation of an effect and thus systematically undermines the internal validity of a study. Bias is systematic insofar that – other than random error – it does not decrease when replicating the study several times or when increasing sample size, the result will always deviate from the true effect[3 4]. Methodologists have defined a large number of individual biases, and the definitions are not always clear[5 6]. One important type of bias is selection bias. It arises when groups are not comparable because of an uneven distribution of patient characteristics and prognostic factors[7-9]. For example, if a study concludes that vegetarianism prolongs life, that may simply be because compared to non-vegetarians, vegetarians tend to smoke less. This is a systematic difference that would introduce selection bias if not controlled for through statistical methods (the effect of not smoking would overshadow the effect of not eating meat). While observational studies are extremely prone to this bias, RCTs can also be concerned when recruiters can guess and alter upcoming treatment allocations, e.g. by lack of allocation concealment leading to a broken randomization[7-9]. Another type of bias is information bias (also called detection or measurement bias) which stems from errors in measurement and determination of exposure and outcome. As these are essential for most statistical analyses, such errors can result in mislead care[3 7]. RCTs are especially prone to this bias when outcome assessors are not blinded[3]. Another major bias in observational studies is confounding. The concept of confounding generally refers to a problem of comparability but it can have different meanings in different scientific fields and eras[8 10]. In epidemiology, a confounder is a factor that influences both the exposure to an intervention and the outcome[7]. For example, cancer stage is a prognostic factor which influences the treatment decision but also the chance of survival (Figure 1).

At the level of study design, RCTs provide methods to control for confounding bias. At the level of analysis, a number of statistical methods are available that aim to control for confounding bias in observational studies: Traditional approaches to control the influence of confounding are, for example, restriction, matching, stratification, multivariate regression, and propensity scores[7 11 12]. These techniques focus on balancing characteristics between comparison groups at baseline. This can be as simple as in the cancer stage example (Figure 1), but confounders cannot always be clearly determined or remain unknown. For example, a study finds a strong association between frequently taking vacations and living longer[13]. It is easy to imagine that stress reduction and increased physical activities may have a positive impact on health and hence on lifespan. A possible confounder could be stress level at work: people with very demanding jobs may not take vacation as often but may have a higher

risk for cardiovascular disease which could shorten their lifespan. It is also possible that people who can afford to go on vacation more frequently have a higher socioeconomic status which is associated with better access to healthcare which can increase their lifespan[13]. The data cannot tell investigators what cause and effect are, or through which factors (or mediators) exposure leads to a specific outcome, or by which other factors it could be influenced. Sometimes investigators can find plausible mechanistic or biologic explanations of exposure as cause for a specific outcome (i.e. causal pathways), e.g. bacteria as cause for many diseases. But still, experts may fail in the attempt to completely understand all underlying factors and base their assumptions on wrong conclusions. For example, many experts criticized the hypothesis that smoking causes cancer and is an important confounder in cancer research[14]. Hence, even when mechanistic explanations are absent, a strong practical effect may still be found. For instance, without understanding why, Ignaz Semmelweis discovered that when he washed his hands with chlorine solution before attending a delivery, more women survived giving birth[15]. In a time where bacteria seemed ridiculous fantasy, this practical approach could still establish cause and effect and saved many lives[15].

To establish cause and effect, RCTs apply a fundamentally different way than observational studies. Instead of trying to statistically control for baseline confounding and risking uneven distributions of patient characteristics between groups, RCTs use chance in their design. By randomly allocating patients to one treatment or the other, RCTs aim at creating equal groups that only differ in the treatment they are intended to receive[2 16]. All known and unknown confounders should, per chance, be divided equally between both groups. If all patients adhered to the protocol, the measured effect would then be the true causal (unconfounded) effect of the treatment[17]. However, perfect adherence is unlikely. As with observational studies, cause and effect can be seen in a mechanistic and a practical way which may both have important aspects in informing treatment choices[18]. Those interested in the mechanistic pathways may now ask how effective the treatment would be if all patients adhered to it, i.e. what is the biological effect. For example, to safely avoid pregnancy, a woman may be more interested in the effect of taking the anti-baby pill at the same time daily (full-adherence) than in taking it with a delay of some hours (non-adherence). This mechanistic question can be answered with a per protocol analysis in which only patients are analyzed that adhered to the treatment protocol. The greater the non-adherence in a trial, the greater the analyzed groups may deviate from the originally randomized ones and confounding is re-introduced. The reason for this is that those who adhere and those who do not may be systematically different, and because adherence may depend on the allocated treatment as well. Conducting a per-protocol analysis then faces the same statistical challenges as observational studies do[19 20]. Those interested in a practical approach may ask how effective the treatment is in general (e.g. the gynecologist cannot know whether or not the patient will actually adhere to taking the anti-baby pill at the same time daily and neither does the patient know this upfront despite her motivation). The practical question is best answered with an intention-to-treat analysis, in which all patients are analyzed according to the groups they were randomly assigned to[18]. The intention-to-treat effect remains unbiased, even if confounding occurs after randomization such as high drop-out rates or treatment switches (i.e. post-randomization confounding)[19]. For example, a physical therapist wants to know if a demanding workout will help patients lose weight compared to a light workout. Because the interventional workout is too demanding, many patients stop working out (drop out) or switch to the light workout. The per-protocol effect may find that the demanding workout resulted in higher weight reduction. However, this effect may be confounded by known and unknown factors: those who adhered to the demanding workout may have had a different body mass index at baseline (this could be statistically controlled for). Their life-situation may have better allowed them

to do the workout regularly (this is less likely to be measured in a trial; factors could be number of people in the household, previous experience with physical programs, motivation, working hours) and also unconscious psychological factors that are not measured or measurable at all may have played a role. Even if all these factors had been known, they would also had to have been adequately measured and then correctly implemented in the model. The unbiased ITT effect, analyzing all patients according to the groups they were allocated to, may not detect a difference between the results of the two interventions. While this does not mean that there is no mechanistic difference between the two interventions[21], it is likely that the less demanding workout would result in a larger average effect than the (in theory) stronger but more demanding workout that only a minority will adhere to. The physical therapist may in future think twice about which patients could benefit from a demanding workout.

Confounding can get even more complex when it varies over time in a longitudinal study. A time-varying confounder is an intermediate variable, i.e. the confounding variable is influenced by previous exposure or changes of the exposure over time. While adjusting for baseline confounders (e.g. prognostic factors that do not change over time, such as sex) reduces bias, adjusting for time-varying confounders may introduce bias when using standard statistical methods[22]. Conventional per-protocol and other standard analyses cannot address this adequately[23]. Marginal structural models (MSMs), a new class of model, can be used to control for time-varying variables and to make causal interpretations[24 25]. They model an alternative scenario, e.g. what would have happened had a patient not taken the treatment but the placebo[25 26]. MSMs are "marginal" because within this framework, the patient population is re-weighted in such way that possible outcomes are independent of possible confounders. For example, if 50 patients receive treatment A and 50 patients receive treatment B, the patients are re-weighted so that each group has 100 patients. This simulates the alternative scenario of what would have happened had the patients who received treatment A actually received treatment B and vice versa. MSMs are "structural" because they attempt to measure a causal effect. To make valid inferences, three main assumptions need to be met: exchangeability, consistency, and positivity[26-28]: Exchangeability means that the groups need to be exchangeable, i.e. there should be no unmeasured confounding (i.e. they are "comparable"). Ideally, this would be the case for the baseline groups of an RCT with perfect randomization and an infinitive large sample. Consistency requires that the exposure is so well defined that variants of it will all lead to the same effect on the outcome. For example, when taking a specific dose of diclofenac for pain relief, there must be no difference in treatment effect when using the products of different pharmaceutical companies. Positivity means that it should be possible for every patient to receive either treatment. For example, positivity is not given if patients are included in the study dataset with an absolute contraindication against the study drug. Overall, MSM is a complex method to plan, conduct, and report but it may give insightful perspectives for study interpretation.

Aims

Confounding is the connecting theme of all projects in this thesis. It may have far-reaching consequences for clinical decision-making[29 30]. Our overall aim was to improve health care decision making by identifying factors that may strengthen or weaken the confidence in evidence used for health care decision-making, and by providing empirical guidance on the utility of MSMs. To achieve this goal, we applied several meta-epidemiological approaches. By using the framework of systematic reviews and meta-analyses, meta-epidemiological research explores the impact of specific study

characteristics on treatment effects and the underlying factors of epidemiological and medical research as a special form of research on research (or meta-research)[3 31].

Objectives

The first doctoral project had the objective to assess whether authors of observational epidemiologic studies considered confounding bias when interpreting their findings. We used a random sample of 120 cohort or case-control studies reporting any exposure-outcome association. The studies were published between 2011 and 2012 by general medical, epidemiological, and specialty journals with the highest impact factors. We evaluated whether the consideration of confounding depended on specific factors, specifically journal types, study types, exposures, journal impact factor and article annual citation rate.

The second doctoral project had the objective to evaluate the agreement between estimated treatment effects of non-randomized studies using causal modelling with marginal structural models and RCTs on the same clinical question. We first included any non-randomized healthcare study that provided an effect from causal modelling with MSM. Then we searched and included RCTs on the same clinical question. In a comparison of the two study designs, we evaluated the direction of treatment effects, effect sizes, and confidence intervals for primary effectiveness outcomes, and the overall absolute deviation. We determined if the effects of the experimental treatment were more or less favorable in non-randomized studies and how the results changed when more RCT evidence was published before the respective non-randomized study.

Intrigued by the emerging use of MSM in RCTs, the third PhD project was a meta-epidemiological analysis with focus on marginal structural models in RCTs. The first objective was to systematically identify and describe situations where MSM had been used to (re-)analyze results from randomized comparisons of medical interventions. Considering all reported results for all available analysis methods within each eligible RCT (e.g. MSM, intention-to-treat, per protocol, as treated), the second objective was to assess the vibration of all effects[32] and the relationship between results of MSM- and intention-to-treat-based analyses.

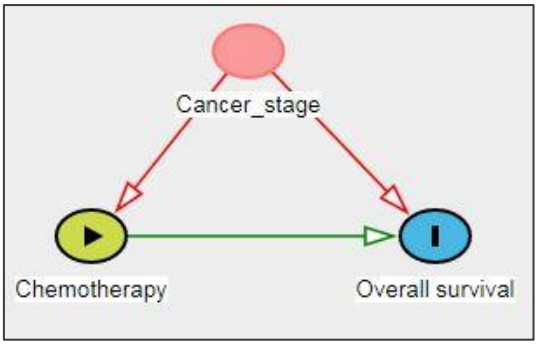

Figure 1 Confounded effect of chemotherapy on overall survival through cancer stage

The arrows denote the proposed causal pathway, i.e. chemotherapy influences the chances of survival in cancer patients, cancer stage influences both chemotherapy and survival and thus confounds the effects of chemotherapy on survival.

Doctoral Manuscripts

I "Interpretation of epidemiologic studies very often lacked adequate consideration of confounding"

Hemkens LG, Ewald H, Naudet F, Ladanie A, Shaw JG, Sajeev G, Ioannidis JPA. Interpretation of epidemiologic studies very often lacked adequate consideration of confounding. Journal of Clinical Epidemiology. 2018;93:94-102.

Status

The manuscript was published in the Journal of Clinical Epidemiology in 2017 ahead of print[33] and finally in January 2018[34].

Abstract

Background and Objective

Confounding bias is a most pervasive threat to validity of observational epidemiologic research. We assessed whether authors of observational epidemiologic studies consider confounding bias when interpreting the findings.

Study Design and Setting

We randomly selected 120 cohort or case–control studies published in 2011 and 2012 by the general medical, epidemiologic, and specialty journals with the highest impact factors. We used Web of Science to assess citation metrics through January 2017.

Results

Sixty-eight studies (56.7%, 95% confidence interval: 47.8–65.5%) mentioned "confounding" in the Abstract or Discussion sections, another 20 (16.7%; 10.0–23.3%) alluded to it, and there was no mention or allusion at all in 32 studies (26.7%; 18.8–34.6%). Authors often acknowledged that for specific confounders, there was no adjustment (34 studies; 28.3%) or deem it possible or likely that confounding affected their main findings (29 studies; 24.2%). However, only two studies (1.7%; 0–4.0%) specifically used the words "caution" or "cautious" for the interpretation because of confounding-related reasons and eventually only four studies (3.3%; 0.1–6.5%) had limitations related to confounding or any other bias in their Conclusions. Studies mentioning that the findings were possibly or likely affected by confounding were more frequently cited than studies with a statement that findings were unlikely affected (median 6.3 vs. 4.0 citations per year, P = 0.04).

Conclusions

Many observational studies lack satisfactory discussion of confounding bias. Even when confounding bias is mentioned, authors are typically confident that it is rather irrelevant to their findings and they rarely call for cautious interpretation. More careful acknowledgment of possible impact of confounding is not associated with lower citation impact.

What is new?

Key findings

- Many highest impact observational studies lack any discussion of confounding bias. Even when mentioned, authors are typically confident that it is rather irrelevant for their findings and they rarely call for cautious interpretation.

What this adds to what was known?

- There is no evidence that acknowledging the potential impact of confounding diminishes citation impact of epidemiological studies.

What is the implication and what should change now?

- There is a need to encourage researchers and to sensitize reviewers and editors to discuss and communicate study limitations introduced by confounding.

Introduction

A confounder may create spurious associations between an exposure and an outcome observed in epidemiologic studies [1]. For example, many more people drinking coffee have lung cancer than people not drinking coffee, but this is because they more often smoke [2]. Many confounders are difficult to pinpoint with certainty, many are entirely unknown, and many others are known, but are still not measured and thus cannot be considered in the analysis of epidemiologic studies. Understanding confounding and separating it from causal effects can be very difficult. For example, even smoking's causal role in cancer, and its potential to confound other observed associations in cancer studies, was not clear across many years of early epidemiologic research [3]. Bias caused by unknown confounders is directly addressable only by randomization, and thus, confounding bias can never be entirely ruled out in nonrandomized studies. Consequently, in the most widely applied framework to assess quality of evidence for healthcare decisions (GRADE), evidence from observational research is initially considered low quality [4].

Because bias due to confounding is a core limitation of observational research, numerous recommendations and statements call for a careful consideration when reporting, discussing, and making conclusions from observational research [[5], [6], [7], [8], [9], [10]]. For example, the Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement, the most widely endorsed guideline for reporting of observational research, prominently emphasizes the discussion of confounding and explicitly states "*It is important not only to identify the sources of bias and confounding that could have affected results, but also to discuss the relative importance of different biases, including the likely direction and magnitude of any potential bias*" and "*due consideration should be given to confounding […]. Authors should also consider residual confounding due to unmeasured variables or imprecise measurement of confounders*" [6].

Despite these recommendations, many investigators might feel that acknowledgment of confounding will cast doubts on their findings. They might prefer to either be silent about this possibility or explicitly discredit the possibility that confounding may have affected their conclusions. Important questions can be asked: Do authors of epidemiologic studies published in major journals acknowledge confounding properly and sufficiently? Does more explicit acknowledgment of confounding as a limitation decrease the subsequent citation impact of their work? To address these questions, here we conducted a meta-epidemiologic survey of observational studies published in high-impact journals. Our primary aim was to assess whether authors of observational epidemiologic studies consider confounding bias when interpreting the findings in the Discussion sections and concluding statements of their articles. Our secondary aim was to determine whether such explicit discussion is associated with lower citation impact.

Methods

Data identification and eligibility

We selected 24 journals with the highest impact factors (Journal Citation Reports 2010): The top eight from the "medicine, general, and internal" category [New England Journal of Medicine, Lancet, JAMA, Annals of Internal Medicine, PLOS Medicine, BMJ, Archives of Internal Medicine (currently JAMA

Internal Medicine), CMAJ], the top eight from the "public, occupational, and environmental health" category (Environmental Health, Epidemiology, International Journal of Epidemiology, American Journal of Epidemiology, Bulletin of the World Health Organization, American Journal of Preventive Medicine, European Journal of Epidemiology, Genetic Epidemiology), and the journal with highest impact factor in each of eight "medical specialty" sub-categories (cardiology and cardiovascular disease, gastroenterology, obstetrics and gynecology, oncology, pediatrics, rheumatology, surgery, urology and nephrology; i.e., Circulation, Gastroenterology, Obstetrics, and Gynecology, Journal of Clinical Oncology, Pediatrics, Annals of Rheumatic Diseases, Annals of Surgery, Journal of the American Society of Nephrology). We did not consider journals focusing exclusively on reviews (e.g., Epidemiologic Reviews) or on basic and/or preclinical research (e.g., Cancer Cell).

We searched MEDLINE for cohort and case–control studies published in these journals in 2011 and 2012 (last search on December 4, 2015; details in Webappendix 1).

The articles retrieved were stratified by journal category. Two independent reviewers (H.E. and F.N.) evaluated randomly selected articles for eligibility until they identified 120 eligible articles (20 per journal type and year; which would allow for standard deviation of <4% for estimated proportions of 75% or 25%). The study flow is shown in Webappendix 2. We included any study clearly described as "cohort study" or "case–control study" (explicitly using these terms) and reporting any exposure–outcome association and thus being theoretically prone to confounding bias. No further eligibility criteria were applied. Any disagreements were resolved by discussion or with a third reviewer (L.G.H.). The random sample included studies published in 22 of the 24 eligible journals (exceptions were Bulletin of the World Health Organization and Genetic Epidemiology), and each journal contributed a median of four studies [interquartile range (IQR) 2–6].

Data extraction

Two independent reviewers (two of L.G.H., H.E., F.N.) extracted the reported study design (i.e., case–control, prospective, retrospective, or unclassified cohort study or nested case–control study; we applied these specific terms to categorize the study design as self-reported by the authors) and categorized the area of research for all pertinent articles. Any disagreements were resolved by discussion or with the third reviewer (L.G.H., H.E., or F.N.).

In addition to manual extractions, two independent reviewers (L.G.H. and H.E.) searched all full-texts automatically (using PDF viewer software) for terms related to propensity scores or marginal structural models anywhere in the articles and they assessed if propensity score–based methods or marginal structural models were used in the studies. There was perfect agreement (100%) between reviewers.

One reviewer (L.G.H.) extracted from Web of Knowledge bibliographic data, specifically the journal's 2010 impact factor and how often the study was cited (Web of Science Core Collection) through January 2, 2017, to calculate an annual citation rate (total citations received per years elapsed since publication).

Evaluation of confounding statements and bias consideration

We systematically evaluated the consideration of confounding bias in the Abstract and Discussion sections of included studies using six standardized prespecified questions (Table 1). We focused on the Abstract and Discussion because these are the sections readers typically focus on the most and from

which they are most likely to draw bottom line conclusions on what the research means and what caveats might exist. We did not evaluate the Introduction, Methods, or Results sections of the publications.

First, we evaluated if the term “confounding” in any form is mentioned at all, regardless of whether it is actually used to discuss the findings of the study or not. We specifically screened Abstract and Discussion sections of the articles for the term “confounding” or variations thereof (Question 1). We also captured any allusions or statements referring to the concept of confounding bias without explicitly using such terms. We also specifically screened the articles for the term “bias” (Question 2) and explicitly perused any mentions of bias for possible relations to confounding. Details with examples are shown in Table 1.

Second, we evaluated if the authors explicitly mention specific potential confounders that were not adjusted for in the analyses (Question 3), or if the authors explicitly discuss whether confounding bias is likely, possible, or unlikely to affect their main findings (Question 4).

Third, we evaluated if confounding bias is considered when interpreting the results or drawing conclusions. Specifically, we evaluated if the authors state that their main results need to be interpreted with caution due to confounding, using the term “caution,” “cautious,” or variants thereof (Question 5). Finally, we specifically screened whether their concluding statements include any limitation or uncertainty related to confounding or bias at all (Question 6). This was evaluated in the section either headed “conclusion,” “summary,” or similar; if such heading did not exist, we evaluated all paragraphs following a concluding statement beginning with, for example, “in conclusion,” or “in summary,” or evaluated the last paragraph of the Discussion.

We developed and pilot tested the operationalization of the questions and iteratively specified the wording of the questions to arrive at detailed extraction instructions. Two reviewers (two of L.G.H., H.E., F.N., A.L.) then assessed all articles independently (unaware of any extractions in the pilot), resolving any disagreements by discussion or with a third reviewer (L.G.H. or H.E.).

Data analysis

In addition to an overall description of the study sample and the statements on confounding, we analyzed whether the consideration of confounding (Questions 1–6) differed between the journal types (general medical vs. epidemiology vs. specialty journal), study types (cohort vs. case–control), exposures (modifiable vs. nonmodifiable), and whether it was associated with journal impact factor and article annual citation rate. We tested differences between continuous variables with the Mann–Whitney U test, differences between categorical data with the Fisher's exact test. Results for continuous measures are medians with IQRs. All analyses were done with Stata 13.1. P values are two tailed.

Results

Evaluated studies

Of the 120 articles, 90 described cohort studies (75%) and 30 case–control studies (25%; Table 2; details in Webappendix 3). Case–control studies were typically published in epidemiologic journals (17 of 30; 56.7%). The 120 studies covered a wide spectrum of medical areas, and there were differences in the areas covered between general medical journals and specialty journals, with pediatrics and oncology being more common in the latter. Most studies (74; 61.7%) analyzed effects of exposures that cannot practically be investigated in experimental studies as they are either not directly modifiable or are harmful (e.g., associations of health outcomes with environmental factors, biomarkers, or demographic characteristics). Effects of potentially modifiable exposures (e.g., drugs, diets, or surgery) were analyzed in 35 studies (29.2%) and were less common in epidemiologic journals. The median impact factor of the 22 journals was 7.9 (IQR, 5.6–13.5) in 2010 and the studies received a median of 5.1 (IQR, 2.5–9.2) annual citations, with clear differences depending on journal type. Of the 120 studies, only six used propensity score methods and one used marginal structural modeling.

Mere mentioning of confounding or bias

Confounding bias was not mentioned or alluded to at all in Abstracts and Discussions of 32 of the 120 studies (26.7%; 95% confidence interval [CI]: 18.8–34.6%; Table 3); in 20 studies (16.7%; 95% CI: 10.0–23.3%), there was some allusion to the concept of confounding indirectly without using this specific term, and 68 of 120 (56.7%; 95% CI: 47.8–65.5%) mentioned the term "confounding" or some same-root variant. The term "bias" was used in 72 of the 120 studies (60%; 95% CI: 51.2–68.8%). Twenty-seven studies (22.5%; 95% CI: 15.0–30.0%) mentioned neither confounding nor bias at all in their Abstracts and Discussions.

Any mention that confounding may affect results

Among the 68 of 120 studies that used the term "confounding" or related terminology, three (2.5%; 95% CI: 0–5.3%) said that it is likely that confounding affects their main findings, 26 (21.7%; 95% CI: 14.3–29.0%) said it is possible, 11 (9.2%; 95% CI: 4.0–14.3%) said it is unlikely, and the remaining 28 did not comment in this regard.

Acknowledgment of unmeasured confounders

Authors of 34 studies (28.3%; 95% CI: 20.3–36.4%) acknowledged that for specific confounders, there was no adjustment, and the reason provided in the majority (28 of 34) was that these confounders had not been measured. Another eight studies mentioned unmeasured confounding in general without specifying the unmeasured confounders.

Cautious interpretation and limitations in conclusions

An explicit statement in the Discussion section (or Abstract) that the interpretation of study results should be made with caution due to possible confounding was made in only 2 of 120 studies (1.7%; 95% CI: 0–4.0%). Specifically, in a study of caffeinated beverage and soda consumption and time to pregnancy, Hatch et al. clearly stated "*We caution that these associations may reflect unmeasured confounding by diet or other lifestyle factors*" [11]. In a study of the association of different biomarkers and risk of type II diabetes, Montonen et al. stated "*Caution is needed when interpreting the results of the analyses on proportion of the association explained. First, the proportion estimates […] may be*

biased if there is unmeasured confounding between the biomarkers and the outcome [References]" [19].

Only 4 of 120 studies (3.3%; 95% CI: 0.1–6.5%) mentioned any limitations related to bias or confounding in their Conclusions.

Of the three studies where the authors' discussion expressed that confounding likely affects their main results, this caution was clearly expressed in the Conclusions in one of the three. Such caution was conveyed in the Conclusion in only 2 of the 26 studies where the authors mentioned possible confounding.

Of the 42 studies where unmeasured confounders were discussed (specifically or in general terms), only one (2.4%) explicitly stated that the interpretation of the results should be made with caution and only four (9.5%) expressed in their Conclusions limitations because of confounding or any other bias.

Overall assessment

The interrater agreement was very high for all assessed questions, ranging from 86.5% to 99.2%. Figure 1 shows the overlap we observed between the different ways of handling and characterizing the potential presence and impact of confounding bias.

Associations with type of journal and impact

The findings were overall the same across the types of journals (Table 3). None of the evaluated aspects of considering confounding bias were associated with journal impact factor or subsequent citation impact, with one exception (Table 4). Studies with a statement that the findings were possibly or likely affected by confounding bias were more frequently cited than those studies with a statement that the findings were unlikely affected (median 6.3 vs. 4.0 citations per year, $P = 0.04$). We found no differences between cohort and case–control studies or between studies evaluating modifiable vs. nonmodifiable exposures (data not shown).

Discussion

Our analysis of 120 randomly selected epidemiologic studies showed that while a narrow majority studies do mention confounding bias to some degree, very few acknowledge that it is a reason for major caution in interpreting the key findings. More than a quarter of the articles completely ignored "confounding" in the Abstract or Discussion sections, and most of them do not even mention the term "bias" in general. Despite the frequent presence and even awareness of specific unmeasured confounders and the often reported possible impact on the main findings, conclusions are almost never made with explicit caution. We found only two cases with explicit statements that cautious interpretation is required because of confounding. Interestingly, in one of them, this caution owing to unmeasured confounding is immediately diluted in the text by stating "*In the present study, we included a large variety of known risk factors as well as of biomarkers, thereby minimizing unmeasured confounding*" [19]. This illustrates the overall impression we gained during our evaluation, that many discussions of confounding in these top journals are superficial and appear to be attempts to negate the importance and impact of confounding in the published work.

We found no indications that this phenomenon is limited to certain areas of research, as findings were similar across types of journals, their impact factors, and study types and topics. Of note, many of the

studies we evaluated were from journals that published the STROBE reporting guidelines in 2007 (i.e., Lancet, Epidemiology, Bulletin of the World Health Organization, BMJ, PLOS Medicine, Annals of Internal Medicine). The observed association of higher study citation numbers with statements acknowledging that confounding bias could exist might be just a chance finding, or be due to confounding. Nevertheless, it suggests that statements acknowledging potential methodological weaknesses have no negative citation impact.

Investigators should not worry that their observational study will be discredited if they acknowledge (as they should) that their work is subject to confounding that might affect their results. Acknowledgment and thorough discussion of the impact of confounding bias may be a marker of researchers with more epidemiologic training being involved in the study, who may have better institutional access to better, larger datasets, and work in larger research teams, all of which may also help explain higher citation rates for articles that explicitly discuss confounding. We did not adjust for any of these potentially explanatory variables in our descriptive analyses as we do not aim to make any causal inferences. If anything, we observed more citations for articles that acknowledged confounding than for those that did not.

The acknowledgment of unmeasured confounding (in accordance to the STROBE reporting guideline) has been systematically assessed in previous empirical work for observational research published in five general medicine journals and five epidemiologic journals (most of them included also in our analysis) for the years 2004–2007 and 2010–2012 [[22], [23]]. Comments on the likelihood of unmeasured confounding were present in 59–85% of the studies, but only 16–32% gave any qualitative statement about the impact on the findings, which agrees well with our overall study results. However, both of these previous empirical studies narrowly evaluated observational research specifically focusing on medical interventions, while we examined the broader landscape of observational investigation within the medical literature, only the minority of which pertained to interventions.

Some limitations of our work deserve closer attention. First, we analyzed only a small sample of the observational study literature. Perhaps, a larger sample may have allowed us to detect small differences between journal types or other factors affecting the consideration of confounding. However, large differences are unlikely to have been missed.

Second, we evaluated studies that were published 4 and 5 years ago, which was necessary for a meaningful analysis of subsequent citation impact. Previous evaluations have found that the introduction of STROBE in 2007, arguably the most influential effort to improve reporting quality, has had only modest impact on reporting quality [[22], [23]]. No new major similar efforts have been launched in the last 5 years; therefore, we have no reason to believe that reporting of observational research would have changed substantially in the last few years.

Third, by only looking at 24 high-impact journals, it is uncertain if our findings are generalizable to the rest of the medical literature. It is quite possible that we may even underestimate the extent to which implications of confounding bias go unaddressed in the medical literature.

We also acknowledge that confounding bias might be seen by some researchers as an inevitable limitation of observational studies that is too well-known to merit discussion. However, as causal interpretations depend on the validity of the implicit assumption of no unmeasured/residual confounding, the implications of bias due to failure of this assumption should be considered. Dealing with confounding bias, understanding its impact (e.g., through qualitative discussion of the magnitude

and direction of bias and more quantitative sensitivity analyses [[24], [25]]), minimizing its influence, and acknowledging the residual uncertainty is an integral core for inference-making in epidemiology. In some situations, authors might not be much interested in causality and expressions about cautious interpretation, for example, when they explore associations for developing diagnostic rules. However, only very few studies in our sample addressed such topics.

Underreporting of limitations may exaggerate conclusions and could sometimes be perceived as sensationalism, overall diminishing trust in research. We found no evidence that considering the possibility of confounding bias diminishes citation impact. This agrees also with recent evaluations of press releases of observational studies showing that cautious interpretations and wide media coverage are well compatible [[26], [27]]. This is reassuring for researchers and may encourage them to discuss and communicate any limitation introduced by confounders in a thorough and determined way and "*not take them as mythical or uncontrollable phantoms that destroy studies*" [28].

Overall, we believe that there is a need to encourage researchers to report more careful and determined considerations of confounding bias and to encourage peer-reviewers, journal editors, and research funders to appreciate this. Many of the journals we analyzed have published the STROBE guideline, and some explicitly refer to them in their Instructions for Authors. Recently, PLOS Medicine intensified the requirements for authors of observational studies, asking that they "*must complete the appropriate reporting checklist not only with page references, but also with sufficient text excerpted from the manuscript to explain how they accomplished all applicable items*" [29]. Our results demonstrate that such activities are well justified. Given that not much has improved over many years, facing the tsunami of big datasets with all their promises, limitations, and risks of spurious findings [30], we believe that more concerted action is needed to improve the appropriate discussion of epidemiologic findings.

Conclusion

Confounding bias is a pervasive threat to the validity of observational epidemiologic research. Inadequate consideration and lack of discussion of implications of confounding bias are very frequent among the highest impact observational studies. Despite reasonable cause for careful discussion and cautious interpretation, authors often convey confidence, without cause or supporting evidence, that confounding bias is largely irrelevant for their findings. We think that such confidence is not justified.

Conflict of interests

All authors have completed the Unified Competing Interest form at www.icmje.org/coi_disclosure.pdf and declare no financial support for this project. L.G.H. is member of the RECORD initiative which aims to improve reporting of observational studies using routinely collected health data. He has no other relationships or activities that could appear to have influenced the submitted work. F.N. has relationships (travel/accommodation expenses covered/reimbursed) with Servier, BMS, Lundbeck, and Janssen who might have an interest in the work submitted in the previous 3 years. He has no other relationships or activities that could appear to have influenced the submitted work. All other authors declare no relationships or activities that could appear to have influenced the submitted work.

Authors' contribution

L.G.H. and J.P.A.I. conceived the study. L.G.H. analyzed the data. All authors interpreted the results. L.G.H. wrote the first draft and all authors made revisions on the article. L.G.H., H.E., A.L., F.N., J.G.S., and G.S. extracted the data. All authors read and approved the final version of the article. L.G.H. and J.P.A.I. are guarantors. All authors had full access to all the data (including statistical reports and tables) in the study and can take responsibility for the integrity of the data and the accuracy of the data analysis.

Funding

This work was supported by a grant of the Laura and John Arnold Foundation to The Meta-Research Innovation Center at Stanford.

Role of the funding source

The funders had no role in design and conduct of the study; collection, management, analysis, and interpretation of the data; and preparation, review, or approval of the article or its submission for publication.

Data sharing

No additional data available.

Ethical approval

Not required for this study.

References

[1] Rothman K, Greenland S, Lash T. Modern epidemiology. 3rd ed. Philadelphia, PA: Lippincott Williams & Wilkins; 2008.

[2] Guertin KA, Freedman ND, Loftfield E, Graubard BI, Caporaso NE, Sinha R. Coffee consumption and incidence of lung cancer in the NIH-AARP Diet and Health Study. Int J Epidemiol 2016;45:929e39.

[3] Cornfield J, Haenszel W, Hammond EC, Lilienfeld AM, Shimkin MB, Wynder EL. Smoking and lung cancer: recent evidence and a discus- sion of some questions. J Natl Cancer Inst 1959;22:173e203.

[4] Balshem H, Helfand M, Schunemann HJ, Oxman AD, Kunz R, Brozek J, et al. GRADE guidelines: 3. Rating the quality of evidence. J Clin Epidemiol 2011;64:401e6.

[5] Morton SC, Costlow MR, Graff JS, Dubois RW. Standards and guide- lines for observational studies: quality is in the eye of the beholder. J Clin Epidemiol 2016;71:3e10.

[6] Vandenbroucke JP, von Elm E, Altman DG, Gotzsche PC, Mulrow CD, Pocock SJ, et al. Strengthening the Reporting of Obser- vational Studies in Epidemiology (STROBE): explanation and elabo- ration. PLoS Med 2007;4(10):e297.

[7] von Elm E, Altman DG, Egger M, Pocock SJ, Gotzsche PC, Vandenbroucke JP, et al. Strengthening the Reporting of Observa- tional Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. BMJ 2007;335:806e8.

[8] Benchimol EI, Smeeth L, Guttmann A, Harron K, Moher D, Petersen I, et al. The REporting of studies Conducted using Observa- tional Routinely-collected health Data (RECORD) statement. PLoS Med 2015;12(10):e1001885.

[9] Berger ML, Mamdani M, Atkins D, Johnson ML. Good research practices for comparative effectiveness research: defining, reporting and interpreting nonrandomized studies of treatment effects using secondary data sources: the ISPOR Good Research Practices for Retrospective Database Analysis Task Force ReportePart I. Value Health 2009;12(8):1044e52.

[10] Developing a Protocol for Observational Comparative Effectiveness Research: A User's Guide. Rockville MD: Agency for Healthcare Research and Quality; 2013.

[11] Hatch EE, Wise LA, Mikkelsen EM, Christensen T, Riis AH, Sorensen HT, et al. Caffeinated beverage and soda consumption and time to pregnancy. Epidemiology 2012;23:393e401.

[12] Press DJ, Sullivan-Halley J, Ursin G, Deapen D, McDonald JA, Strom BL, et al. Breast cancer risk and ovariectomy, hysterectomy, and tubal sterilization in the women's contraceptive and reproductive experiences study. Am J Epidemiol 2011;173:38e47.

[13] Jackson LA, Yu O, Nelson JC, Dominguez C, Peterson D, Baxter R, et al. Injection site and risk of medically attended local reactions to acellular pertussis vaccine. Pediatrics 2011;127(3):e581e7.

[14] O'Reilly CE, Jaron P, Ochieng B, Nyaguara A, Tate JE, Parsons MB, et al. Risk factors for death among children less than 5 years old hospi- talized with diarrhea in rural western Kenya, 2005-2007: a cohort study. PLoS Med 2012;9(7):e1001256.

[15] Ferguson LP, Durward A, Tibby SM. Relationship between arterial partial oxygen pressure after resuscitation from cardiac arrest and mortality in children. Circulation 2012;126:335e42.

[16] Niederkrotenthaler T, Rasmussen F, Mittendorfer-Rutz E. Perinatal conditions and parental age at birth as risk markers for subsequent suicide attempt and suicide: a population based case-control study. Eur J Epidemiol 2012;27(9):729e38.

[17] Coupland C, Dhiman P, Morriss R, Arthur A, Barton G, Hippisley- Cox J. Antidepressant use and risk of adverse outcomes in older people: population based cohort study. BMJ 2011;343:d4551.

[18] Cook AG, deVos AJ, Pereira G, Jardine A, Weinstein P. Use of a total traffic count metric to investigate the impact of roadways on asthma severity: a case-control study. Environ Health 2011;10:52.

[19] Montonen J, Drogan D, Joost HG, et al. Estimation of the contribu- tion of biomarkers of different metabolic pathways to risk of type 2 diabetes. Eur J Epidemiol 2011;26(1):29e38.

[20] Mazumdar M, Bellinger DC, Gregas M, et al. Low-level environ- mental lead exposure in childhood and adult intellectual function: a follow-up study. Environ Health 2011;10:24.

[21] Lacson E Jr, Xu J, Suri RS, et al. Survival with three-times weekly in- center nocturnal versus conventional hemodialysis. J Am Soc Neph- rol 2012;23(4):687e95.

[22] Groenwold RH, Van Deursen AM, Hoes AW, Hak E. Poor quality of reporting confounding bias in observational intervention studies: a systematic review. Ann Epidemiol 2008;18:746e51.

[23] Pouwels KB, Widyakusuma NN, Groenwold RH, Hak E. Quality of reporting of confounding remained suboptimal after the STROBE guideline. J Clin Epidemiol 2016;69:217e24.

[24] Vanderweele TJ, Arah OA. Bias formulas for sensitivity analysis of unmeasured confounding for general outcomes, treatments, and con- founders. Epidemiology 2011;22:42e52.

[25] Lash TL, Fox MP, MacLehose RF, et al. Good practices for quantita- tive bias analysis. Int J Epidemiol 2014;43:1969e85.

[26] Sumner P, Vivian-Griffi S, Boivin J, et al. The association between exaggeration in health related science news and aca- demic press releases: retrospective observational study. BMJ 2014;349:g7015.

[27] Sumner P, Vivian-Griffiths S, Boivin J, Williams A, Bott L, Adams R, et al. Exaggerations and caveats in press releases and health-related science news. PLoS One 2016;11:e0168217.

[28] Vandenbroucke JP. The history of confounding. Soz Praventivmed 2002;47(4):216e24.

[29] Plos Medicine Editors. Observational studies: getting clear about transparency. PLoS Med 2014;11(8):e1001711.

[30] Hemkens LG, Contopoulos-Ioannidis DG, Ioannidis JP. Routinely collected data and comparative effectiveness evidence: promises and limitations. CMAJ 2016;188(8):E158e64.

Tables

Table 1: Assessment of consideration of confounding bias in Abstracts and Discussions

1.	Do the authors mention confounding using explicitly the terms "confounder(s)," "confounding," "confound," or do they allude to it without using those terms, or is confounding not considered at all? Examples for "yes": *"We caution that these associations may reflect unmeasured confounding by diet or other lifestyle factors"* [11]. Example for "alluded": *"Another potential limitation is our inability to control for age at menopause among women having a hysterectomy before natural menopause; for these women, age at menopause is unknown"* [12]. *"When we included the characteristics we could define in multivariable models the association of arm injection site with a significantly higher risk of medically attended local reactions persisted, but it is possible that bias may have influenced the findings"* [13].
2.	Do the authors mention bias using explicitly the term "bias"? Example for "yes": *"Where available, we relied on HIV diagnosis based on clinical features, which may be subject to biases in assessing the factors contributing to diarrheal disease among participants since HIV infection at early stages may have been missed and not all data were routinely captured"* [14].
3.	Do the authors mention specific confounders that have not been adjusted for? (If yes, what were the reasons? If not, were there unspecified unmeasured confounders without specifically stating which ones?) Example for "yes": *"We were unable to adjust for additional confounding variables with a known association with mortality (for example, blood glucose and postarrest pH) that were not collected as part of the PICANet data set"* [15].
4.	Do the authors state that their main findings are likely, possibly, or unlikely affected by residual confounding? Example for "yes, likely": *"Therefore, some residual confounding with parental psychopathology seems likely"* [16].

	Example for "yes, possibly": *"However, although we adjusted for severity of the initial diagnosis of depression, we could use only a crude measure as we did not have a validated depression severity score. We cannot therefore exclude the possible effect of residual confounding on our results"* [17]. Example for "yes, unlikely": *"Minimal differences were observed between the crude and adjusted odds ratios for the exposure variable, suggesting that SEIFA and ethnicity were unlikely to be major confounders in this analysis"* [18].
5.	Do the authors state that their main findings need to be interpreted with caution due to confounding? We answered this question with "yes" in cases with a clear statement that cautious interpretation is required because of confounding. Example for "yes": *"Caution is needed when interpreting the results of the analyses on proportion of the association explained. First, the proportion estimates, decomposed from the total effect by adjusting for other biomarkers, may be biased if there is unmeasured confounding between the biomarkers and the outcome [Reference]. In the present study, we included a large variety of known risk factors as well as of biomarkers, thereby minimizing unmeasured confounding"* [19].
6.	Do the authors call for caution or indicate limitations or uncertainty due to possible confounding or other bias in their conclusions? Example for "yes": "*We caution that these associations may reflect unmeasured confounding by diet or other lifestyle factors"* [11]. *"Given the small sample size, however, the potentially confounding effects of maternal IQ cannot be excluded and should be evaluated in a larger study"* [20]. *"In summary, notwithstanding the possibility of residual selection bias, patients who […]"* [21].

Table 2: Characteristics of studies

Study characteristics	Total, no. (%)	Journal category			*P*-value
		General medicine, no. (%)	Epidemiology, no. (%)	Medical specialties, no. (%)	
Number of studies	120 (100)	40 (100)	40 (100)	40 (100)	–
Study design					<0.01
Case–control	22 (18.3)	3 (7.5)	13 (32.5)	6 (15.0)	–
Nested case–control study[a]	8 (6.7)	2 (5.0)	4 (10.0)	2 (5.0)	–
Cohort study, prospective	48 (40.0)	19 (47.5)	17 (42.5)	12 (30.0)	–
Cohort study, retrospective	25 (20.8)	8 (20.0)	2 (5.0)	15 (37.5)	–
Cohort study, unclassified	17 (14.2)	8 (20.0)	4 (10.0)	5 (12.5)	–
Area of disease or condition					<0.01
Cardiology, CVD	12 (10.0)	5 (12.5)	5 (12.5)	2 (5.0)	–
Obstetrics and gynecology	16 (13.3)	6 (15.0)	8 (20.0)	2 (5.0)	–
Oncology	16 (13.3)	0 (0.0)	7 (17.5)	9 (22.5)	–
Pediatrics	27 (22.5)	6 (15.0)	7 (17.5)	14 (35.0)	–
Other	49 (40.8)	23 (57.5)	13 (32.5)	13 (32.5)	–
Type of exposure					<0.01
Pathogens	4 (3.3)	2 (5.0)	0 (0.0)	2 (5.0)	
Genetics	5 (4.2)	0 (0.0)	2 (5.0)	3 (7.5)	
Diet	5 (4.2)	3 (7.5)	2 (5.0)	0 (0.0)	
Surgery	6 (5.0)	1 (2.5)	1 (2.5)	4 (10.0)	
Demographic characteristics	7 (5.8)	1 (2.5)	5 (12.5)	1 (2.5)	
Comorbidities	9 (7.5)	4 (10.0)	3 (7.5)	2 (5.0)	
Diagnostics/prediction rules	12 (10.0)	5 (12.5)	1 (2.5)	6 (15.0)	
Environmental factors	13 (10.8)	1 (2.5)	11 (27.5)	1 (2.5)	
Biomarkers	14 (11.7)	3 (7.5)	6 (15.0)	5 (12.5)	
Drug treatment	14 (11.7)	6 (15.0)	1 (2.5)	7 (17.5)	
Nonmodifiable, other, or multiple	10 (8.3)	4 (10.0)	2 (5.0)	4 (10.0)	
Modifiable, other, or multiple	17 (14.2)	8 (20.0)	4 (10.0)	5 (12.5)	
Modifiable and nonmodifiable	4 (3.3)	2 (5.0)	2 (5.0)	0 (0.0)	
Citation impact					
IF 2010 (median, IQR)	7.9 (5.6–13.5)	14.5 (13.5–30.0)	5.7 (4.5–5.7)	7.9 (5.4–12.0)	
(range), n = 120	2.5–53.5	9.0–53.5	2.5–5.9	4.4–19.0	
Citations/year (median, IQR)	5.0 (2.6–9.8)	9.1 (4.8–19.7)	3.7 (2.3–5.1)	5.1 (2.5–9.2)	
(range), n = 120	0.2–66.7	1.3–66.7	0.2–11.1	0.7–33.6	

Abbreviations: CVD, cardiovascular disease; IF, impact factor; IQR, interquartile range.

[a] Including two case–cohort studies.

Table 3: Statements on confounding

Question	Journal category				*P*-value interrater agreement
	Total, no. (%)	General medicine, no. (%)	Epidemiology, no. (%)	Medical specialties, no. (%)	
Total	120 (100)	40 (100)	40 (100)	40 (100)	
1. "Confounding" mentioned in Abstract or Discussion?					0.33 88.2%
Yes, specific term	68 (56.7)	24 (60.0)	26 (65.0)	18 (45.0)	
Alluded	20 (16.7)	6 (15.0)	7 (17.5)	7 (17.5)	
No	32 (26.7)	10 (25.0)	7 (17.5)	15 (37.5)	
2. Term "Bias" used in Abstract or Discussion?					0.30 93.6%
Yes	72 (60.0)	27 (67.5)	25 (62.5)	20 (50.0)	
No	48 (40.0)	13 (32.5)	15 (37.5)	20 (50.0)	
3. Specific nonadjusted confounders acknowledged?					0.50 89.8%
Yes	34 (28.3)	11 (27.5)	14 (35.0)	9 (22.5)	
...because not measured	28 (82.4)	11 (100)	12 (85.7)	5 (55.6)	0.039
...because of other reasons	4 (11.8)	0 (0)	2 (14.3)	2 (22.2)	
...no reasons given	2 (5.9)	0 (0)	0 (0)	2 (22.2)	
No	86 (71.7)[a]	29 (72.5)	26 (65.0)	31 (77.5)	
4. Any mention that findings may be affected by confounding?					0.39 86.5%[b]
Likely	3 (2.5)	1 (2.5)	1 (2.5)	1 (2.5)	
Possibly	26 (21.7)	10 (25.0)	8 (20.0)	8 (20.0)	
Unlikely	11 (9.2)	3 (7.5)	7 (17.5)	1 (2.5)	
No statement	80 (66.7)	26 (65.0)	24 (60.0)	30 (75.0)	
5. Cautious interpretation needed?					0.33 99.2%
Yes	2 (1.7)	0 (0)	2 (5.0)	0 (0)	
No	118 (98.3)	40 (100)	38 (95.0)	40 (100)	
6. Conclusions include any limitations?					>0.99 98.3%
Yes	4 (3.3)	1 (2.5)	2 (5.0)	1 (2.5)	
No	116 (96.7)	39 (97.5)	38 (95.0)	39 (97.5)	

[a] In 8 of the 86 studies, unmeasured confounding was mentioned, but no specific confounder stated.
[b] Interrater agreement calculated only for the 40 studies making a statement.

Table 4: Citation impact

Question	No. of studies	Journal if (median, IQR)	*P*-value	Citations per year (median, IQR)	*P*-value
1. "Confounding" mentioned in Abstract or Discussion?					
Yes	68	9.0 (5.7–14.4)	0.46	5.4 (2.6–9.5)	0.69
No or allude	52	6.7 (5.4–13.5)		4.6 (2.6–10.8)	
2. Term "Bias" used in Abstract or Discussion?					
Yes	72	8.7 (5.7–14.4)	0.24	5.2 (2.7–8.8)	0.79
No	48	6.7 (5.4–13.5)		4.9 (2.5–12.4)	
3. Specific nonadjusted confounders acknowledged?					
Yes	34	9.0 (5.7–13.5)	0.73	5.6 (2.2–9.3)	0.72
No	86	6.7 (5.4–13.5)		4.8 (2.7–10.2)	
4. Any mention that findings may be affected by confounding?					
Possibly or likely	29	13.5 (5.7–14.4)	0.07	6.4 (4.7–10.2)	0.04
Unlikely	11	5.7 (2.5–13.5)		4.0 (2.2–5.1)	
5. Cautious interpretation needed?					
Yes	2	5.2 (4.5–5.9)	0.30	2.4 (2.2–2.5)	0.15
No	118	8.3 (5.7–13.5)		5.1 (2.7–10.0)	
6. Conclusions include any limitations?					
Yes	4	7.1 (4.2–10.9)	0.59	9.7 (5.3–21.9)	0.28
No	116	7.9 (5.6–14.0)		4.9 (2.6–9.5)	

IF: Journal Citation Reports 2010 Impact Factor.

Figures

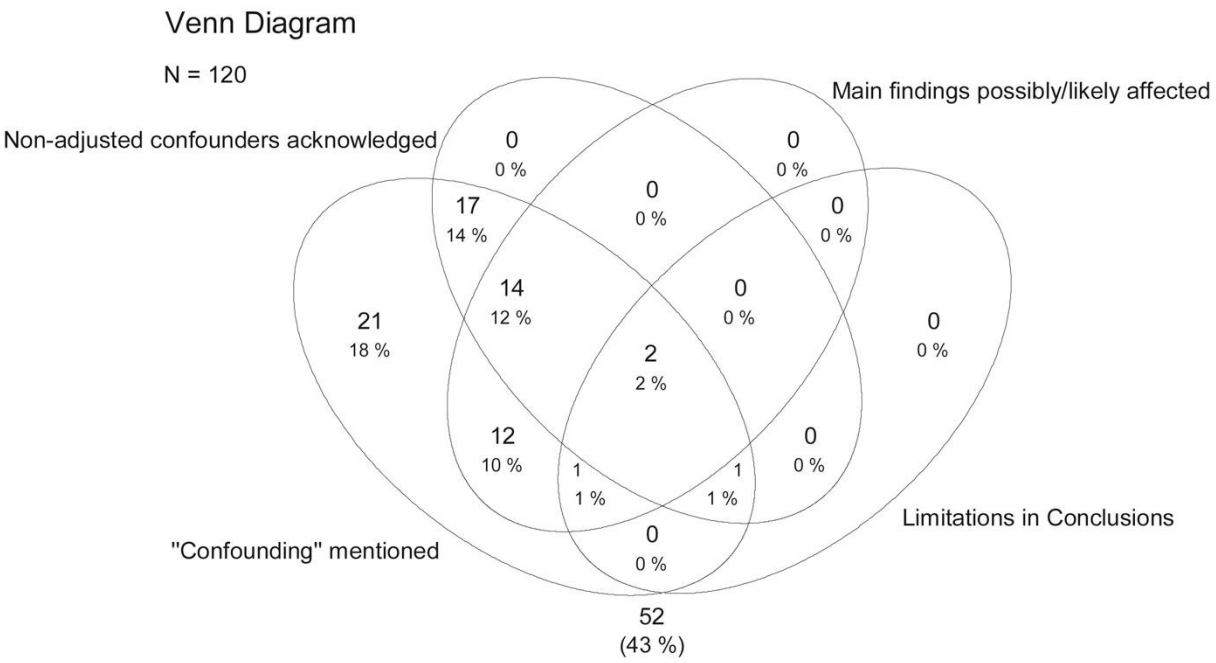

Figure 1: Venn diagram on different aspects of consideration of confounding bias in discussions of epidemiologic research. Each ellipsoid area corresponds to one aspect of consideration of confounding bias. The numbers indicate the number of studies sharing the characteristics in the overlapping areas, for example, there are 14 epidemiologic studies (12% of 120) in which “confounding” is mentioned in the Abstract or Discussion, the authors deem the main findings possibly or likely affected by confounding, and nonadjusted confounders are acknowledged, but there are no limitations in the Conclusions related to confounding or any bias. Fifty-two studies are not covered by any of the areas. The percentages do not correspond to the size of circular areas.

Webappendix

Supplementary data related to this article can be found at https://doi.org/10.1016/j.jclinepi.2017.09.013.

II "Impact of Marginal Structural Models as enhanced confounder control methods in non-randomized comparative effectiveness: a meta-epidemiologic study"

Ewald H, Ioannidis JPA, Ladanie A, Mc Cord K, Bucher HC, Hemkens LG

Status

The manuscript was under peer review by the BMJ but not considered for publication. We plan to submit it to another high-impact journal in April 2018.

Abstract

Objectives

To evaluate the agreement of estimated treatment effects between non-randomized studies using causal modelling (with marginal structural models) and randomized controlled trials (RCTs) on the same clinical question.

Design

Meta-epidemiological study.

Data sources

PubMed, Scopus and citations of key references searched up to April 2017.

Methods

Any non-randomized study using marginal structural models for causal modelling providing an effect estimate on any healthcare outcome of any treatment was eligible. We systematically sought RCTs on the same clinical question and compared the direction of treatment effects, effect sizes, and confidence intervals for primary effectiveness outcomes between study designs. We evaluated the overall absolute deviation between study designs and we assessed if non-randomized studies found more or less favorable effects for the experimental treatment. We used the ratio of odds ratios (ROR; the summary odds ratio of trials divided by the reported estimate from non-randomized studies) for each clinical question and then combined the RORs using random-effects meta-analysis. Meta-regression was used to assess whether the agreement between study designs is associated with previous knowledge of RCT-effects.

Results

The main analysis included 19 non-randomized studies with 1039570 patients and 141 RCTs with 120669 patients. Non-randomized studies indicated treatment effect estimates in the opposite direction from RCTs for 8 clinical questions (42%), and their 95% confidence interval did not include the RCT estimate in 9 clinical questions (47%). The effect estimates deviated systematically by 1.29-fold (summary absolute deviation OR 1.29; 95% confidence interval 1.12 to 1.48). Overall, causal modelling studies tended to show more favorable results for the experimental treatment (summary

ROR 1.14; 95% confidence interval 0.93 to 1.41), in particular when they clearly focused on healthcare decision making and clinical interpretation not on statistical methodology (16 studies, summary ROR 1.34; 95% confidence interval 1.03 to 1.75), and when more RCT evidence was previously available (p=0.037).

Conclusions

Non-randomized studies using causal modelling with marginal structural models may give different answers than RCTs evaluating the same clinical question. Caution continues to be required when non-randomized “real world” evidence is used to guide health care decisions.

What is known on this topic:

- Many health care stakeholders call for use of non-randomized “real world” evidence for decision-making
- Non-randomized studies are prone to bias due to confounding
- Causal modelling with marginal structural models may theoretically overcome such biases under the critical assumption that all confounders are known, correctly measured, and precisely included in the models

What does it add:

- Non-randomized “real world” studies using causal modelling may give different answers than randomized controlled trials with effects that typically deviate and sometimes show stronger benefits of experimental treatments
- Caution remains crucial when non-randomized evidence is used to guide health care decisions – even when causal modelling techniques were applied and especially when no evidence from randomized controlled trials exists

Introduction

Randomized controlled trials (RCTs) are able to determine the causal effects of choosing between two or more treatment options for an intervention.[1 2] However, RCTs are not available for many important healthcare questions. Since healthcare decision makers then have to rely on non-randomized observational studies as the only evidence, attempts are made to increase the reliability of these studies.[2 3] To assess the causal effect of a treatment choice, it would be inadequate to simply compare outcomes between patients selecting different treatments. The reason is that in the "real world", the choice is not random. It often depends on specific factors, such as comorbidities or personal values and preferences of patients and care providers, that are also associated with future health outcomes (confounding by indication). Many health care stakeholders call for use of non-randomized "real world" evidence and its evaluation in the context of health care decision-making – even for drug approval as in recent US legislation.[4-7] Although only randomization can directly eliminate confounding, several methods are used in observational studies in the attempt to address this issue, for example multivariate regression analysis or methods using propensity score.[8 9]

We have recently shown that studies based on routinely collected health data analyzed with propensity score methods may substantially overestimate treatment benefits and may provide very different answers than RCTs on the same clinical question.[10] However, these methods balance confounders at baseline without addressing confounders arising during follow-up when patients switch or stop treatments because of unsatisfactory results.[11] Novel approaches have been proposed aiming to eliminate this time-dependent confounding and to measure causal effects of treatments in such situations.[12] The most frequently used method involves marginal structural models (MSM). MSMs are a relatively new class of statistical models that are increasingly applied for causal inference in non-randomized studies over the last two decades.[11 13 14] The underlying essential inherent limitation of all non-randomized study designs remains the same,[15] that is the assumption that all relevant confounders are known, measured, and correctly integrated in the analyses.

There is no meta-epidemiologic analysis evaluating if non-randomized "real world" evidence using causal modelling with MSM [16] can reliably estimate effects as found in RCTs, which would be a prerequisite for reliably using this evidence for healthcare decision-making. We aimed to evaluate the agreement of treatment effects estimated by non-randomized studies using MSM for causal modelling with effects of RCTs investigating the same clinical question in a comprehensive meta-epidemiological study.

Methods

We analyzed non-randomized studies that evaluated any healthcare intervention with causal modelling using MSM (MSM-studies). We systematically compared the findings from such studies for various clinical questions with the findings from RCTs on the same clinical question.

Identification of non-randomized studies using marginal structural models

We included any observational, non-randomized study that (1) provided at least one effect estimate of any binary health outcome clearly based on the use of MSM, (2) mentioned the intervention, the evaluated population, and any result (not necessarily MSM-based) in the abstract, (3) investigated a clearly defined treatment and comparator . We applied no language restrictions.

We screened the bibliography and citations of 12 key references[11 13 16-25] on MSM using Web of Science (last search 23 June 2014) to identify relevant MSM-studies. We then used the search results for the development of a PubMed search strategy (including terms related to "marginal structural models"; last search 15 October 2014; Webappendix) to identify further pertinent articles. Two reviewers (HE, LGH) independently screened all titles, abstracts and full-texts, and solved any disagreement through discussion.

One reviewer (HE) then identified the clinical questions (following the PICO scheme[26], i.e. the composite of population, intervention, comparison, and outcome) addressed in each eligible MSM-study and a second reviewer (LGH) verified this.

Identification of corresponding randomized trial evidence

First, we searched for systematic reviews that reported treatment effects of RCTs that assessed the same clinical question as the MSM-study. One reviewer (HE) developed search strategies for all clinical questions using a standardized search building approach and applying standard filters for systematic reviews and meta-analyses.[27] Potentially relevant RCTs cited in the MSM-study were used as benchmark to maximize the sensitivity of the search, i.e. we adapted the search until we retrieved all these potentially relevant studies. All search strategies were checked by a second reviewer (LGH). One reviewer (HE) then searched PubMed (last search 21 April 2016), screened all titles, abstracts, and full-texts. She recorded any article that she deemed possibly eligible; another reviewer (LGH) verified eligibility of all included or possibly relevant studies (LGH). Another reviewer evaluated a random 10% sample of all clinical questions that were not deemed included or possibly relevant (LGH or AL). The interrater agreement was perfect (100%).

Second, we hand-searched the reference lists of eligible MSM-study publications for any mentioned or cited RCTs on the same clinical question.

Third, we screened all studies citing eligible MSM-study publications for eligible RCTs using SCOPUS (last search 28 March 2017).

Fourth, whenever we found eligible RCTs that were not covered by a systematic review discovered in the first step (this was the case for 3 clinical questions), we updated our search for systematic reviews and also did a complete search on PubMed for RCTs on that clinical question using the PubMed standard filter for RCTs (last search 12 April 2017).

One reviewer (HE) screened all reference lists, titles and abstracts. All potentially relevant full-texts were independently screened by two reviewers (HE, LGH) who solved any disagreement through discussion.

Data Extraction

From each MSM-study, we extracted each clinical question with corresponding MSM-based treatment effect. We extracted RCT-based treatment effects from available systematic reviews of RCTs (we used individual trial results if available for all but three clinical questions[28-30] where we extracted the summary effect estimates with confidence intervals from meta-analyses in the systematic review). We used RCT publications when there was no systematic review. Information on the conduct of intention-to-treat analyses, the number of patients missing (i.e. lost-to-follow-up, withdrawn, discontinued or dropped out), and the number of patients who switched their allocated treatment were extracted from

RCT publications. For information on risk of bias of the RCTs, we extracted the assessments from available systematic reviews, otherwise we used the original RCT data applying Cochrane standards.[27] We asked systematic review authors for bibliographic information if it was not clear to us which RCTs were included in their meta-analyses. One reviewer extracted and assessed all data (HE) and another verified them (KM).

Statistical analysis

For the consistent reflection of treatment effects across all clinical questions from the MSM-studies and RCTs, we inverted the effect estimates and confidence intervals where necessary so that the results represent effects of experimental treatments versus control treatments for an unfavorable outcome. The first treatment comparator always reflects the experimental group (i.e. newer, less established, treatment of interest) and the second comparator the control group (i.e. standard of care, placebo, older treatment). Where this classification was not clear, we consulted experts in the field. Whenever the outcome was favorable, we inverted the reported study effect to reflect the complementary unfavorable outcome (e.g. death instead of survival). For the RCTs, we used the same outcome and direction of comparison as we had used in the MSM-studies.

When an MSM-study evaluated several outcomes for the same treatment comparison in the same population, we selected the (all-cause) mortality outcome for the main analysis because mortality is of high relevance for decision makers, rather unambiguously measured, and probably less prone to data accuracy problems in routinely collected data sources (mortality was the primary endpoint in 2 of 3 MSM-studies that were concerned[31 32]). When there were multiple effect estimates reported for the same outcome (i.e. different results for one clinical question, derived from various MSMs), we used the effect estimate mentioned first.

We assumed that reported relative risks or hazard ratios in MSM-studies are close approximations of ORs (which is reasonable when the event rate is low,[33] as was in our sample with a median event rate of the intervention and control groups of 5.5%, IQR 1% to 12%).

Treatment effects of multiple RCTs on the same clinical question were combined to obtain one summary OR per clinical question with random effects model meta-analyses using the DerSimonian and Laird method.

First, we examined how often MSM-studies and RCTs differed in the direction of the point estimates; how often the MSM-study and the respective RCTs had the same direction in the presence of statistical significance (their 95% confidence intervals excluded the null effect); how often the RCTs' summary estimate was not included in the 95% confidence interval of the respective MSM-study; and how often the treatment effects from MSM-studies and RCTs differed beyond chance.

Second, we quantified the absolute deviation between OR estimates from MSM-studies and estimates from corresponding RCTs for each clinical question (calculated as absolute difference between the RCT-based and MSM-based log(OR)s and back transformed on the OR-scale). This provides an estimate of deviation between study designs independent from assumptions about effect directions, experimental and control treatments. The absolute deviation estimate and the confidence interval limits are positive by definition and reported as x-fold deviation on the OR scale. For example, the absolute deviation would be 1.25-fold when one study design finds an OR = 1 and the other design finds an OR = 0.8 or an OR = 1.25.

Third, we assessed the directions of deviations to explore if effects found in MSM-studies for experimental treatments would be more or less favorable than in RCTs. For this, we used the ratio of odds ratio (ROR) approach, i.e. the division of the summary OR of RCTs by the MSM-based treatment effect (after log-transformation).[34] A ROR for experimental treatment effects above 1 indicates that the MSM-study measured a more favorable result of the experimental treatment benefit than the RCTs, a ROR below 1 indicates less favorable treatment benefits found in MSM-studies.

The absolute deviation ORs and the RORs for experimental treatment effects of all clinical questions were each synthesized with random effects models (DerSimonian and Laird) to obtain overarching summaries of the relationship of effects in non-randomized studies using causal modelling versus corresponding randomized trial evidence.

Fourth, we explored if there were indications to assume that results of MSM-studies would be different when they are done without pre-existing randomized evidence on treatment effects[35]. Because almost all clinical questions in our analysis had pre-existing randomized evidence when the causal model results were published, we could not apply the cleanest approach by including only MSM-studies without pre-existing trials.[35] Instead we conducted a meta-regression in which we estimated the association between the ROR per clinical question and the amount of randomized evidence available at the time of the publication of the MSM-study. The amount of evidence was expressed by the inverse of the variance (i.e. by the weight in the random-effects meta-analysis of all the RCTs with publication dates before the MSM-study). We also conducted sensitivity analyses excluding all RCTs with a publication date before the MSM-study or excluding all RCTs published after the MSM-study.

We conducted further sensitivity analyses (whenever data were available on at least 5 clinical questions) excluding 3 MSM-studies focusing on statistical methodology of causal modelling without having any clinical interpretations for healthcare decision making in their conclusions[16 36 37]; including only all-cause mortality effects; including only non-mortality effects (here, in case of several pertinent estimates, we selected the most precise one, i.e. the one with smallest confidence interval); excluding clinical questions with active treatment controls; excluding clinical questions without active treatment controls; excluding RCTs with either high risk of bias in any of the bias domains, and/or missing data of more than 10%, and/or where more than 10% of the patients switched their allocated treatment; and using a fixed-effect model to combine data from RCTs.

We used Stata 14.2 (Stata Corp, College Station, TX, USA) for all analyses. We used the 'metan'-command for meta-analyses, the "metareg"-command for meta-regression, and the 'heterogi'-command for the confidence intervals of I^2 [38]. P-values are two tailed.

Patient involvement

No patients were involved in this research.

Results

Studies evaluated

The literature search for MSM-studies resulted in 3916 references; we screened titles and abstracts and obtained 646 articles in full-text. We found 98 eligible articles on causal modelling studies using

MSM. In search for corresponding RCT-evidence, we screened 9926 references. Overall, we found 168 RCTs corresponding to 25 clinical questions addressed in 19 of the 98 MSM-studies (Figure 1, Table 1).

Three of the 19 MSM-studies[31 32 39] evaluated more than one clinical question, i.e. they assessed several outcomes (all including all-cause mortality which we used for the main analysis) in the same population and treatment comparison (Table 1). Of the 19 MSM-studies included in the main analysis, 14 (76%) had all-cause mortality as outcome. Most MSM-studies were related to HIV (6/19, 32%) or nephrology (8/19, 42%) and explored drug treatment comparisons (15/19, 79%). This was similar to the topics in all identified MSM-studies (HIV 43/98, 44%; nephrology 19/98, 19%; 52/98, 53% on drugs). The observational data were mostly from registries (8/19, 42%) and electronic medical records (6/19, 32%) routinely collected in various countries.

MSM-studies were published between 2001 and 2014 and RCTs between 1968 and 2017. In total, 133 of 168 RCTs were clearly published before the date of publication of the MSM-study (79%). For 11 (58%) of the 19 MSM-studies, all corresponding RCTs were published before the MSM-study.

Agreement of treatment effects

Overall 1039570 patients (median of 9939 patients per clinical question, IQR 990 to 51037) were evaluated from causal modelling studies and 141 corresponding RCTs were identified with overall 120669 patients (median of 229 patients per RCT, IQR 68 to 946) for the 19 clinical questions included in the main analysis.

The direction of effect estimates was in opposite directions in 8 of 19 clinical questions (42%). In 9 of 19 (47%) clinical questions, the 95% confidence interval of the MSM-studies did not include the RCT treatment effect estimate (Table 2). In 9 clinical questions (47%), the two designs differed on whether or not they showed statistically significant differences between the compared treatments (significant differences found only by MSM-studies in 8 cases, and only by RCTs in 1 case), and in 4 of these 9 clinical questions, the point estimates of the two designs were in opposite directions. The 95% confidence intervals for one or both designs were typically large, thus the estimates differed beyond chance in only 1 of the 19 clinical questions (i.e. the 95% confidence interval of the ROR excluded 1).

The absolute deviation of effect sizes between study designs across all 19 clinical questions was 1.29-fold (summary OR 1.29 95% confidence interval 1.12 to 1.48). When we summarized the RORs, we found a non-significant trend for more favorable results for the experimental treatments estimated by non-randomized studies with causal modelling than by RCTs (summary ROR 1.14; 95% confidence interval 0.93 to 1.41; Table 2; Figure 2). The non-randomized studies tended to have more favorable results for the experimental treatment when there was more evidence from RCTs previously available (meta-regression p = 0.037).

In all sensitivity analyses (Table 2), the absolute deviations of effect sizes between study designs were quite similar, ranging from 1.18-fold to 1.72-fold. Causal modelling studies with a clear focus on guiding healthcare decisions and making clinical interpretations systematically overestimated benefits of experimental treatments (summary ROR 1.34; 95% confidence interval 1.03 to 1.75). Causal modelling studies underestimated treatment benefits when only trials that were published after the MSM-study were included (summary ROR 0.73; 95% confidence interval 0.48 to 1.11).

Discussion

Principal findings

This comprehensive meta-epidemiological analysis found that results of non-randomized "real world" studies using causal modelling with MSM frequently do not agree with RCTs evaluating the same clinical question. Although our results need to be interpreted with caution due to the limited sample size, the principal finding was that MSM-study results deviate from RCTs 1.29-fold and show greater benefits of experimental treatments than RCTs. The direction of treatment effects estimated with causal models is often opposite to that estimated in RCTs (43%), and their 95% confidence intervals provide little guidance to tell what RCTs on the same topic would show. When causal modeling studies focus on healthcare decision making, they systematically and significantly exaggerated clinical benefits.

Comparison with other studies

To our knowledge, this is the first systematic empirical study on how treatment effect estimates derived from non-randomized studies using causal modelling analyses agree with RCTs.[40] Previous evaluations of the credibility of results from "real world" non-randomized evidence analyzed with causal modeling were based on comparisons with trials in very specific and selected situations.[11 18 41] This study evaluated all MSM-studies identified with a large, reproducible, highly sensitive search strategy. Overall, our findings are very similar to recent analyses of "real world" evidence based on propensity scores.[35] Assumptions that reported results from observational data analyses are substantially influenced by knowledge of expected treatment effects from RCTs receive further support.[35]

Limitations

This study has some limitations that merit closer attention. First, although we found 98 MSM-studies, our final sample covered only 19 of them. For the other studies and their clinical questions, we did not find any relevant randomized trial. We sought for close agreement of the research questions in both study designs to minimize bias. There may have been more clinical questions that our search did not capture as it was widely based on published systematic reviews. However, conducting more than 300 de-novo, high-quality systematic reviews for all clinical questions with sensitive search strategies to identify, assess, and extract any individual trials would not have been feasible. However, our citation-based searches rarely detected randomized evidence and also the large number of empty reviews (i.e. reviews that searched for pertinent trials on a clinical question but did not find any) indicates that for numerous clinical questions there is indeed no trial. Overall, the clinical fields covered by our sample, with many questions in the field of HIV, nephrology, and on drug effects, are very similar to the overall use of MSM in observational research.

Second, several MSM-study authors have known the results from pre-existing RCTs on the explored topic, and some explicitly used them to improve their modelling, illustrate methodological aspects or to compare treatment effects between the study designs. Most of the MSM-studies were published after the corresponding RCTs and MSM-study authors cited a corresponding RCT or systematic review for almost half of the clinical questions (45%). This may have created publication or reporting bias as it may affect the decision to report or publish results that are in strong disagreement with pre-existing trial evidence. A clean approach avoiding such influences and resembling the natural situation of decision-making in the absence of clinical trial evidence and comparison with subsequently generated knowledge[35] was not possible. These mechanisms may have resulted in a better agreement between

MSM-studies with RCTs in situations with existing trial evidence. This is supported by the results from the meta-regression showing an association of prior RCT knowledge with observational estimates and also by the sensitivity analysis excluding the three MSM-studies focusing on statistical methodology.

Third, from the trials, we preferred intention-to-treat analyses as they theoretically allow causal estimation of effects and are most robust against biases. This approach may lead to smaller treatment effect estimates in the trials especially with high rates of treatment switching and missing data and no active control.[42 43] However, excluding clinical questions with no active controls resulted in even larger absolute differences between the study designs. Excluding trials with high rates of treatment switching or missing data and high risk of bias also showed larger absolute deviation of effects. However, the reporting quality was often insufficient and for many trials the switching rates, attrition and risk of bias were unclear. Conceptual divergences between starting and adhering to the treatment instead of the intention to treat[43 44] were difficult to assess as the reporting was often also unclear in the MSM-studies, which is a known problem.[41] Although such theoretical issues may explain some of the observed differences between both designs, more and better reported information would be needed for further evaluation. Clinical conclusions of the MSM-studies never highlighted this issue. Any interpretation in this regard should be made with caution.

Fourth, we decided not to assess the quality or risk of bias of the MSM-studies as this is not straightforward in observational studies, particularly in those using MSM. The recently published ROBINS-I tool ("Risk Of Bias In Non-randomised Studies of Interventions") is the only suitable instrument we are aware of but it only asks general questions on time-varying confounding.[44] The critical assumption for unbiased causal inference in observational research is that all relevant confounders are known, correctly measured, and precisely integrated in the analyses. The absence of unmeasured confounding has been explicitly stated by many of the study authors but typically without clear justifications or calls for caution. It would be entirely subjective to judge whether confounders or other known or unknown factors were truly irrelevant, which would be a prerequisite of a low risk of bias. It is generally unlikely that confounding would be completely eliminated, for example, when patient preferences and values determine treatment choice or adherence and outcomes.

Overall, it is important to note that our evaluation neither intended to nor can compare the theoretical advantages of causal modelling versus well conducted randomized trials. Instead, we aimed to compare results of published evidence on treatment effects derived from studies reporting the use of such approaches as health care decision makers such as clinicians would use them as guidance. The main focus of three of the MSM-studies was not the support of healthcare decision making but statistical methods. They were done under more extraordinary circumstances with selected topics and data sources, much larger and clearer existing randomized evidence, and overall more exemplary characteristics than studies conducted by researchers who focus more on guiding every day healthcare decisions. When we only included "real world" evidence clearly focusing on guiding healthcare decisions, the difference between study designs was larger with a clear overestimation of clinical benefits beyond chance in the MSM-studies compared to RCTs.

Conclusion

The use of non-randomized "real world" evidence in the context of health care decision-making needs to be seen with caution. Causal modelling may have theoretical advantages and offer valuable insights but only when certain assumptions hold true. The critical assumptions that all factors which may bias

the results are known, correctly measured, and precisely included in the models may be unrealistic. Available randomized evidence on treatment effects may influence reported results of non-randomized studies which adds further uncertainty to situations in which decisions must be made without any trial evidence. When randomized trials are unavailable and clearly unfeasible, decision makers may rely on such analyses, but they should be very aware that benefits may often be smaller and more uncertain than non-randomized "real world" evidence suggests – despite causal modelling with marginal structural models. To better protect patients from hazardous consequences of misguided decisions from non-randomized analyses, efforts should be undertaken to generate, whenever possible, randomized "real world evidence", such as simple, large, pragmatic trials that address critically important clinical questions and guide health care more reliably.

Acknowledgements

The authors thank Kübra Özoglu, University of Basel, for her administrative assistance and support with literature management, Michael Koller, MD, University of Basel, for assistance in the interpretion of study data from nephrology, and authors of original systematic reviews for their responses to information requests.

Data sharing

No additional data available.

Declaration of competing interests

All authors have completed the Unified Competing Interest form at www.icmje.org/coi_disclosure.pdf. All authors declare no financial relationships with any organization that might have an interest in the submitted work in the previous three years and no other relationships or activities that could appear to have influenced the submitted work.

Authors' contribution

HE, LGH, HCB conceived the study with input on the study design by JPAI. HE, LGH, KM, AL extracted the data. HE and LGH analyzed the data. HE, LGH, JPAI, HCB interpreted the results. HE wrote the first draft and all authors made revisions on the manuscript. All authors read and approved the final version of the paper. LGH acquired funding for this study. HE and LGH are the guarantors.

Funding

This work was supported by the Gottfried and Julia Bangerter-Rhyner-Foundation. The Meta-Research Innovation Center at Stanford is funded by a grant by the Laura and John Arnold Foundation. The Basel Institute of Clinical Epidemiology and Biostatistics is supported by Stiftung Institut für klinische Epidemiologie.

Role of the funding source

The funders had no role in design and conduct of the study; collection, management, analysis, and interpretation of the data; and preparation, review, or approval of the manuscript or its submission for publication.

Transparency declaration

The Corresponding Author affirms that the manuscript is an honest, accurate, and transparent account of the study being reported; that no important aspects of the study have been omitted; and that any discrepancies from the study as planned have been explained.

Ethical approval

Not required, this article does not contain any personal medical information about any identifiable living individuals.

References

1. Stel VS, Dekker FW, Zoccali C, et al. Instrumental variable analysis. Nephrology, dialysis, transplantation : official publication of the European Dialysis and Transplant Association - European Renal Association 2013;28(7):1694-9.

2. Armstrong K. Methods in comparative effectiveness research. Journal of clinical oncology : official journal of the American Society of Clinical Oncology 2012;30(34):4208-14.

3. Luce BR, Kramer JM, Goodman SN, et al. Rethinking randomized clinical trials for comparative effectiveness research: the need for transformational change. Annals of internal medicine 2009;151(3):206-9.

4. Sherman RE, Davies KM, Robb MA, et al. Accelerating development of scientific evidence for medical products within the existing US regulatory framework. Nat Rev Drug Discov 2017;16(5):297-98.

5. Sherman RE, Anderson SA, Dal Pan GJ, et al. Real-World Evidence — What Is It and What Can It Tell Us? New England Journal of Medicine 2016;375(23):2293-97.

6. Franklin JM, Schneeweiss S. When and How Can Real World Data Analyses Substitute for Randomized Controlled Trials? Clinical Pharmacology & Therapeutics 2017;102(6):924-33.

7. Wall KM, Kilembe W, Vwalika B, et al. Hormonal Contraceptive Use Among HIV-Positive Women and HIV Transmission Risk to Male Partners, Zambia, 1994-2012. J Infect Dis 2016;214(7):1063-71.

8. Kyriacou DN, Lewis RJ. Confounding by Indication in Clinical Research. Jama 2016;316(17):1818-19.

9. Braga LHP, Farrokhyar F, Bhandari M. Practical Tips for Surgical Research: Confounding: What is it and how do we deal with it? Canadian Journal of Surgery 2012;55(2):132-38.

10. Hemkens LG, Contopoulos-Ioannidis DG, Ioannidis JPA. Agreement of treatment effects for mortality from routinely collected data and subsequent randomized trials: meta-epidemiological survey. BMJ (Clinical research ed) 2016;352:i493.

11. Suarez D, Borras R, Basagana X. Differences between marginal structural models and conventional models in their exposure effect estimates: a systematic review. Epidemiology (Cambridge, Mass) 2011;22(4):586-8.

12. Mansournia MA, Higgins JP, Sterne JA, et al. Biases in Randomized Trials: A Conversation Between Trialists and Epidemiologists. Epidemiology (Cambridge, Mass) 2017;28(1):54-59.

13. Robins JM, Hernan MA, Brumback B. Marginal structural models and causal inference in epidemiology. Epidemiology (Cambridge, Mass) 2000;11(5):550-60.

14. Snowden JM, Rose S, Mortimer KM. Implementation of G-computation on a simulated data set: demonstration of a causal inference technique. American journal of epidemiology 2011;173(7):731-8.

15. Goodman SN, Schneeweiss S, Baiocchi M. Using Design Thinking to Differentiate Useful From Misleading Evidence in Observational Research. Jama 2017;317(7):705-07.

16. Hernán MA, Brumback B, Robins JM. Marginal Structural Models to Estimate the Joint Causal Effect of Nonrandomized Treatments. Journal of the American Statistical Association 2001;96(454):440-48.

17. Cole SR, Hernan MA. Constructing inverse probability weights for marginal structural models. American journal of epidemiology 2008;168(6):656-64.

18. Hernan MA, Brumback B, Robins JM. Marginal structural models to estimate the causal effect of zidovudine on the survival of HIV-positive men. Epidemiology (Cambridge, Mass) 2000;11(5):561-70.

19. Hernan MA, Robins JM. Estimating causal effects from epidemiological data. J Epidemiol Community Health 2006;60(7):578-86.

20. Robins J. Marginal structural models. 1997 Proceedings of the Section on Bayesian Statistical Science. Alexandria: American Statistical Association, 1998;1-10.

21. Robins JM. Correction for non-compliance in equivalence trials. Statistics in medicine 1998;17(3):269-302; discussion 87-9.

22. Robins JM. Association, causation, and marginal structural models. Synthese 1999;121(1-2):151-79.

23. Robins JM. Marginal Structural Models versus Structural nested Models as Tools for Causal inference. In: Halloran ME, Berry D, eds. Statistical Models in Epidemiology, the Environment, and Clinical Trials. New York, NY: Springer New York, 2000:95-133.

24. Robins JM, Greenland S, Hu F-C. Estimation of the Causal Effect of a Time-Varying Exposure on the Marginal Mean of a Repeated Binary Outcome. Journal of the American Statistical Association 1999;94(447):687-700.

25. VanderWeele TJ. Marginal structural models for the estimation of direct and indirect effects. Epidemiology (Cambridge, Mass) 2009;20(1):18-26.

26. Richardson WS, Wilson MC, Nishikawa J, et al. The well-built clinical question: a key to evidence-based decisions. ACP J Club 1995;123(3):A12-3.

27. Higgins JPT, Green S. Cochrane Handbook for Systematic Reviews of Interventions Version 5.1.0 [updated March 2011]. The Cochrane Collaboration 2011;Available from www.cochrane-handbook.org.

28. Dubicka B, Hadley S, Roberts C. Suicidal behaviour in youths with depression treated with new-generation antidepressants: meta-analysis. Br J Psychiatry 2006;189:393-8.

29. Ioannidis JP, Cappelleri JC, Skolnik PR, et al. A meta-analysis of the relative efficacy and toxicity of Pneumocystis carinii prophylactic regimens. Arch Intern Med 1996;156(2):177-88.

30. Knight SR, Russell NK, Barcena L, et al. Mycophenolate mofetil decreases acute rejection and may improve graft survival in renal transplant recipients when compared with azathioprine: a systematic review. Transplantation 2009;87(6):785-94.

31. Cain LE, Logan R, Robins JM, et al. When to initiate combined antiretroviral therapy to reduce mortality and AIDS-defining illness in HIV-infected persons in developed countries: an observational study. Annals of internal medicine 2011;154(8):509-15.

32. Hernandez D, Muriel A, Abraira V, et al. Renin-angiotensin system blockade and kidney transplantation: a longitudinal cohort study. Nephrology, dialysis, transplantation : official publication of the European Dialysis and Transplant Association - European Renal Association 2012;27(1):417-22.

33. Davies HTO, Crombie IK, Tavakoli M. When can odds ratios mislead? BMJ : British Medical Journal 1998;316(7136):989-91.

34. Sterne JA, Juni P, Schulz KF, et al. Statistical methods for assessing the influence of study characteristics on treatment effects in 'meta-epidemiological' research. Statistics in medicine 2002;21(11):1513-24.

35. Hemkens LG, Contopoulos-Ioannidis DG, Ioannidis JP. Agreement of treatment effects for mortality from routinely collected data and subsequent randomized trials: meta-epidemiological survey. BMJ (Clinical research ed) 2016;352:i493.

36. Danaei G, García Rodríguez LA, Cantero OF, et al. Observational data for comparative effectiveness research: an emulation of randomised trials to estimate the effect of statins on primary prevention of coronary heart disease. Statistical methods in medical research 2013;22(1):70-96.

37. Delaney JA, Daskalopoulou SS, Suissa S. Traditional versus marginal structural models to estimate the effectiveness of beta-blocker use on mortality after myocardial infarction. Pharmacoepidemiology and drug safety 2009;18(1):1-6.

38. Higgins JP, Thompson SG. Quantifying heterogeneity in a meta-analysis. Statistics in medicine 2002;21(11):1539-58.

39. de Beaudrap P, Etard JF, Gueye FN, et al. Long-term efficacy and tolerance of efavirenz- and nevirapine-containing regimens in adult HIV type 1 Senegalese patients. AIDS Res Hum Retroviruses 2008;24(6):753-60.

40. Anglemyer A, Horvath HT, Bero L. Healthcare outcomes assessed with observational study designs compared with those assessed in randomized trials. The Cochrane database of systematic reviews 2014;4:Mr000034.

41. Yang S, Eaton CB, Lu J, et al. Application of marginal structural models in pharmacoepidemiologic studies: a systematic review. Pharmacoepidemiology and drug safety 2014;23(6):560-71.

42. Abraha I, Cherubini A, Cozzolino F, et al. Deviation from intention to treat analysis in randomised trials and treatment effect estimates: meta-epidemiological study. BMJ (Clinical research ed) 2015;350:h2445.

43. Hernán MA, Hernández-Díaz S. Beyond the intention to treat in comparative effectiveness research. Clinical trials (London, England) 2012;9(1):48-55.

44. Sterne JA, Hernan MA, Reeves BC, et al. ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions. BMJ (Clinical research ed) 2016;355:i4919.

45. Gibbons RD, Coca Perraillon M, Hur K, et al. Antidepressant treatment and suicide attempts and self-inflicted injury in children and adolescents. Pharmacoepidemiology and drug safety 2015;24(2):208-14.

46. Hocqueloux L, Choisy P, Le Moal G, et al. Pharmacologic boosting of atazanavir in maintenance HIV-1 therapy: the COREYA propensity-score adjusted study. PLoS One 2012;7(11):e49289.

47. Kainz A, Heinze G, Korbely R, et al. Mycophenolate mofetil use is associated with prolonged graft survival after kidney transplantation. Transplantation 2009;88(9):1095-100.

48. Khanal N, Marshall MR, Ma TM, et al. Comparison of outcomes by modality for critically ill patients requiring renal replacement therapy: a single-centre cohort study adjusting for time-varying illness severity and modality exposure. Anaesth Intensive Care 2012;40(2):260-8.

49. Lukowsky LR, Mehrotra R, Kheifets L, et al. Comparing mortality of peritoneal and hemodialysis patients in the first 2 years of dialysis therapy: a marginal structural model analysis. Clin J Am Soc Nephrol 2013;8(4):619-28.

50. Marshall MR, Hawley CM, Kerr PG, et al. Home Hemodialysis and Mortality Risk in Australian and New Zealand Populations. American Journal of Kidney Diseases;58(5):782-93.

51. Mehrotra R, Chiu YW, Kalantar-Zadeh K, et al. Similar outcomes with hemodialysis and peritoneal dialysis in patients with end-stage renal disease. Arch Intern Med 2011;171(2):110-8.

52. Petersen ML, Wang Y, van der Laan MJ, et al. Virologic efficacy of boosted double versus boosted single protease inhibitor therapy. AIDS 2007;21(12):1547-54.

53. Sterne JA, Hernan MA, Ledergerber B, et al. Long-term effectiveness of potent antiretroviral therapy in preventing AIDS and death: a prospective cohort study. Lancet 2005;366(9483):378-84.

54. Teng M, Wolf M, Ofsthun MN, et al. Activated injectable vitamin D and hemodialysis survival: a historical cohort study. J Am Soc Nephrol 2005;16(4):1115-25.

55. Tentori F, Albert JM, Young EW, et al. The survival advantage for haemodialysis patients taking vitamin D is questioned: findings from the Dialysis Outcomes and Practice Patterns Study. Nephrology, dialysis, transplantation : official publication of the European Dialysis and Transplant Association - European Renal Association 2009;24(3):963-72.

56. Tiihonen J, Lonnqvist J, Wahlbeck K, et al. 11-year follow-up of mortality in patients with schizophrenia: a population-based cohort study (FIN11 study). Lancet 2009;374(9690):620-7.

57. Wiesbauer F, Heinze G, Mitterbauer C, et al. Statin use is associated with prolonged survival of renal transplant recipients. J Am Soc Nephrol 2008;19(11):2211-8.

Tables

Table 1: Description of clinical questions evaluated in non-randomized studies using marginal structural models and corresponding randomized trial evidence

MSM-study	Total no. of patients in MSM-study / corresponding RCTs	Clinical question*	Country Data source Collection period
Danaei 2013[36]	74806 / 70902	Statins in cardiovascular disease prevention on fatal or non-fatal myocardial infarction	UK The Health Improvement Network (THIN) database electronic medical records 2000-2006
de Beaudrap 2008[39]	217 / 1067	Efavirenz vs. nevirapine in HIV-infected patients receiving 2 nucleoside reverse transcriptase inhibitors **	Senegal "Initiative Sénégalaise d'Accès aux Medicaments Anti-rétroviraux" ISAARV prospective cohort 1998-2002
Delaney 2009[37]	9939 / 21847	Beta-blocker after myocardial infarction	UK General Practice Research Database (GPRD) records 2002-2004
Gibbons 2014[45]	55284 / 2741	Antidepressants in children on suicide attempts and self-inflicted injuries	USA Medical claim database: MarketScan 2004-2009
Hernan 2001[16]	2168 / 823	Pneumocystis jiroveci pneumonia-prophylaxis in HIV-infected patients	USA Multicenter AIDS Cohort Study (MACS) dataset 1984-1991
Hernandez 2012[32]	990 / 51	Angiotensin-converting enzyme inhibitors and/or angiotensin II receptor blockers in renal transplant recipients ***	Spain Regional transplant center database medical records 1996-2005

HIV Causal Collaboration 2011[31]	20971 / 4685	Antiretroviral therapy initiation based on CD4 cell count	Multinational**** Cohorts 1996-2009
Hocqueloux 2012[46]	352 / 887	Ritonavir-boosted vs. unboosted atazanavir in HIV-infected patients receiving 2 nucleoside reverse transcriptase inhibitors on virological failure	France COREYA study (COhort with REYAtaz); retrospective cohort of electronic database 2004-2011
Kainz 2009[47]	1219 / 2647	Mycophenolate mofetil vs. azathioprine in renal transplant patients	Austria Austrian Dialysis and Transplant Registry OEDTR, Vienna Kidney Biopsy Registry and subsets of the EUROTRANSPLANT databases 1996-2005
Khanal 2012[48]	146 / 561	Prolonged intermittent renal replacement therapy vs. continuous renal replacement therapy in acute kidney injury or acute-on-chronic kidney disease	New Zealand Middlemore Hospital clinical records 2002-2008
Lukowsky 2013[49]	23718 / 38	Peritoneal dialysis vs. hemodialysis	USA Database records from the US Renal Data System (USRDS) and DaVita 2001-2006
Marshall 2011[50]	26016 / 38	Peritoneal dialysis vs. conventional facility hemodialysis	Australia and New Zealand Dialysis and Transplant Registry 1996-2007
Mehrotra 2011[51]	684426 / 38	Peritoneal dialysis vs. hemodialysis	USA National registry for all patients with ESRD 2002-2004
Petersen 2007[52]	988 / 130	Boosted double vs. boosted single protease inhibitor in HIV infected patients	USA Kaiser Permanente Medical Center patients 1998-2005
Sterne 2005[53]	1276 / 8389	Triple vs. dual antiretroviral therapy in HIV-infected patients on AIDS or death	Switzerland Swiss HIV Cohort Study records 1996-1999
Teng 2005[54]	51037 / 112	Injectable vitamin D in patients on chronic hemodialysis	USA Medical records of dialysis facilities

			1996-2002
Tentori 2009[55]	38066 / 597	Vitamin D in patients on dialysis	France, Germany, Italy, Japan, Spain, UK, USA, Australia, Belgium, Canada, New Zealand and Sweden Dialysis Outcomes and Practice Patterns Study (DOPPS) medical records 1996-2007
Tiihonen 2009[56]	66881 / 2356	Antipsychotics in schizophrenia	Finland Nationwide registers 1996-2006
Wiesbauer 2008[57]	2041 / 2760	Statins in renal transplant recipients	Austria Austrian Dialysis and Transplant Registry 1990-2005

*) The comparator was usual care/no treatment and the outcome all-cause mortality if not specified otherwise

**) Further outcomes evaluated were virological success (2250 patients included in corresponding RCTs), virological failure (n=1571), treatment discontinuation for any reason (n=1648), new CDC-C event (n=1215)

***) Further outcome evaluated was death-censored graft failure (114 patients included in corresponding RCTs)

****) UK CHIC (United Kingdom Collaborative HIV Cohort), ATHENA (AIDS Therapy Evaluation Netherlands), FHDH-ANRS CO4 (French Hospital Database on HIV—Agence Nationale de Recherches sur le SIDA), SHCS (Swiss HIV Cohort Study), PISCIS (Proyecto para la Informatización del Seguimiento Clinico-epidemiológico de la Infección por HIV y SIDA [Spain]), CoRIS (Cohorte de la Red de Investigación en SIDA), US VACS-VC (United States Veterans Aging Cohort Study–Virtual Cohort), UK Register of HIV Se-roconverters, ANRS PRIMO and ANRS SEROCO (Agence Nationale de Recherches sur le SIDA, France), GEMES (Grupo Español Multicéntrico para el Estudio de Seroconvertores-Haemophilia)

Abbreviations: AIDS: Acquired immune deficiency syndrome; CDC-C event: Centers for Disease Control and Prevention classification system for HIV-infection, category C: severely symptomatic; HIV: human immunodeficiency virus; MSM-Study: non-randomized study using marginal structural models

Table 2: Agreement of treatment effects from non-randomized studies using marginal structural models and randomized trial evidence

Analysis	Absolute deviation* (95% CI) Heterogeneity (95% CI)	Summary ROR (95% CI) Heterogeneity (95% CI)	No. of clinical questions n (percent)				
			Total	Point estimates of MSM-study and RCTs have opposite directions	Point estimates of MSM-study and RCTs have same direction and exclude the null effect	95% CI of MSM-study exclude the RCTs' point estimate	95% CI of ROR excludes the null effect
Main analysis	1.29 (1.12 to 1.48) 0% (0 to 43%)	1.14 (0.93 to 1.41) 30% (0 to 59%)	19	8** (42%)	2 (11%)	9 (47%)	1 (5%)
Sensitivity analyses							
Fixed-effect model	1.27 (1.12 to 1.44) 0% (0 to 43%)	1.11 (0.98 to 1.26) 31% (0 to 59%)	19	8** (42%)	2 (11%)	9 (47%)	2 (11%)
MSM-studies with a clear focus on guiding treatment decisions	1.61 (1.28 to 2.02) 0% (0 to 45%)	1.34 (1.03 to 1.75) 19% (0 to 55%)	16	7** (43%)	1 (6%)	9 (56%)	1 (6%)
Only RCTs published before the MSM-study	1.29 (1.11 to 1.49) 0% (0 to 45%)	1.26 (1.00 to 1.59) 30% (0 to 60%)	16	8** (50%)	2 (13%)	8 (50%)	1 (6%)
Only RCTs published after the MSM-study	1.39 (0.92 to 2.09) 0% (0 to 61%)	0.73 (0.48 to 1.11 0% (0 to 61%)	6	3** (50%)	0 (0%)	2 (25%)	0 (0%)

Mortality outcomes only	1.26 (1.08 to 1.48) 0% (0 to 47%)	1.16 (0.88 to 1.54) 36% (0 to 65%)	14	8** (57%)	1 (7%)	9 (64%)	1 (7%)
Non-mortality outcomes only	1.42 (1.17 to 1.73) 0% (0 to 56%)	0.98 (0.70 to 1.37) 58% (0 to 79%)	8	0 (0%)	2 (25%)	2 (25%)	1 (13%)
Active comparators only	1.72 (1.29 to 2.30) 0% (0 to 53%)	1.31 (0.86 to 1.99) 44% (0 to 72%)	10	4 (40%)	1 (10%)	4 (40%)	1 (10%)
Non-active comparators	1.18 (1.01 to 1.38) 0% (0 to 54%)	1.04 (0.88 to 1.23) 3% (0 to 56%)	9	4** (44%)	1 (11%)	5 (56%)	0 (0%)
Excluding RCTs with high risk of bias, >10% missing outcome data or >10% treatment switch	1.43 (1.16 to 1.75) 0% (0 to 56%)	0.97 (0.70 to 1.34) 50% (0 to 76%)	8	3 (38%)	1 (13%)	4 (50%)	1 (13%)

* Summary OR

** Including one study where the MSM-effect estimate is below 1 and the RCT-effect estimate is exactly 1

CI: Confidence Interval; RCT: randomized controlled trial; MSM-Study: non-randomized study using marginal structural models

Figures

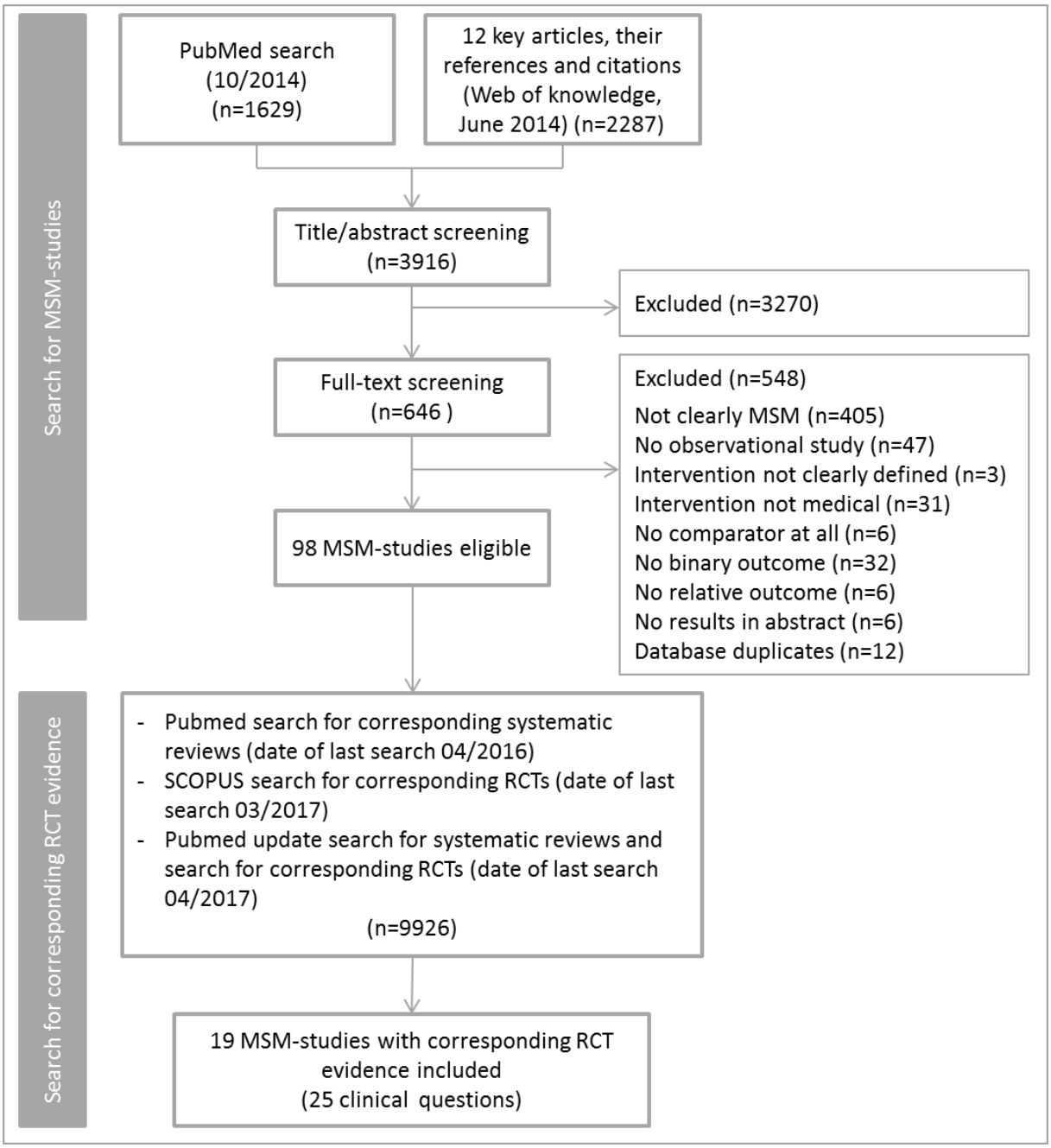

Figure 1: Study selection process

MSM-Studies: non-randomized studies using marginal structural models; RCT: randomized controlled trial

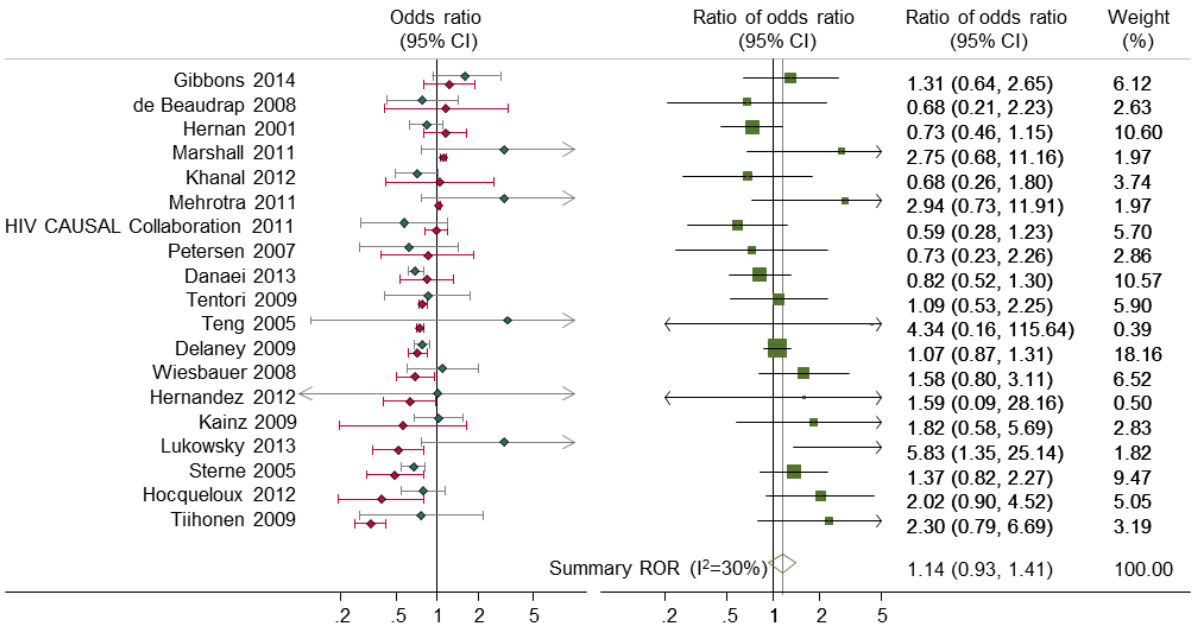

Figure 2 Treatment effects estimated with non-randomized studies using marginal structural models and randomized trial evidence

Left panel shows effect estimates (odds ratios) and 95% confidence intervals (CIs) of healthcare interventions on different outcomes reported in 19 non-randomized studies using marginal structural models for causal modelling (lower graphs; red diamonds indicate treatment effect estimates and lines indicate the 95% CIs) and in randomized controlled trials investigating the same clinical question (upper graphs; blue diamonds and lines).

Right panel shows for each clinical question the ratio of outcome effects reported in non-randomized studies versus randomized trial evidence (as relative odds ratios, ROR, with green squares as effect estimate and lines as 95% CIs). The combined summary ROR (random-effects meta-analysis of RORs) across all 19 clinical questions is shown as diamond. Values greater than 1 indicate more favorable results for the experimental treatment by non-randomized studies using causal modelling.

Webappendix

Search strategy for non-randomized studies using marginal structural models

Search terms	Hits
((IPTW[tiab] OR "inverse probability"[tiab]) OR (marginal[tiab] AND structur*[tiab] AND model*[tiab]) OR (marginal[tiab] AND "models, structural"[mh])) NOT (ANIMALS[mh] NOT HUMANS[mh])	1629

Search date: 15 October 2014

Database and Interface: PubMed

III "Treatment effects from marginal structural models in randomized clinical trials: meta-epidemiological analysis"

Ewald H, Speich B, Ladanie A, Bucher HC, Ioannidis JPA, Hemkens LG

Status

The manuscript is ready for submission. Submission to Trials is planned for March 2018.

Abstract

Objective

Marginal structural models (MSMs) are increasingly used to estimate causal effects of treatments, typically in non-randomized studies. They are one method to deal with time-dependent confounding arising from non-adherence. We determined how MSMs are used in randomized clinical trials (RCTs) and compared their results against those obtained with intention-to-treat (ITT) or other analyses.

Study Design and Setting

We systematically searched PubMed, Scopus, citations of key references, and Clinicaltrials.gov till May 2017. We included RCTs reporting effects of any health care interventions based on MSMs and on at least one other (ITT, as treated, per protocol) analysis.

Results

We included 12 RCTs published between 2002 and 2016 including a median of 1972 patients (interquartile range (IQR) 870 to 17006) reporting 138 analyses for 24 treatment comparisons (median 6 analyses, IQR 3 to 7 per comparison). On average, the largest reported effect estimate from any analysis was 1.19-fold (median on a relative risk scale; IQR 1.13 to 1.34) larger than the smallest reported effect estimate for the same comparison. All MSM and ITT-based results were in the same direction and had overlapping 95% confidence intervals, and in 71% (12 of 17 with CI) they also agreed on the presence or not of nominal statistical significance. For 13 of 20 comparisons (65%), MSM-based relative risks were more extreme (deviated more from the null) than ITT (p=0.18) (median, 1.11-fold (IQR 0.99 to 1.22)). The MSM- and ITT-based estimated relative risks differed on median 1.12-fold (IQR 1.02 to 1.22).

Conclusion

MSMs typically provided similar results as ITT and other available analyses. Some of the differences in effect estimates or nominal significance may nevertheless become important in clinical decision making, but also require utmost attention of possible selective reporting bias.

Introduction

Randomized clinical trials (RCTs) are usually the best way to measure causal effects of treatments. RCTs allow to measure the causal effect of being assigned to a treatment using the intention-to-treat (ITT) approach [1], and they may allow to estimate the effect of initiating and continuously being adherent to the treatment using the "per protocol" (PP) or "as treated" (AT) approach [2].

Trials may be designed to answer clinical questions about the practical consequences of deciding to initiate a treatment, such as prescribing an antibiotic, beginning a life-style intervention or a treatment which requires good adherence. Ideally, the decision to initiate a treatment (the "intention to treat") is followed by an actual start of the treatment with close adherence to the protocol. Randomization allows to measure effects of such decisions without confounding baseline variables. Such trials focusing on health care decision making are often pragmatic or practical [3]. Explanatory or mechanistic trials aim to better understand the underlying causal pathways of the decisions, such as biological mechanisms of treatment effects, or effects specifically in highly-motivated patients with high probability to be compliant to study protocols [3]. For such research questions, PP effects (evaluating only patients who adhere to the study protocol) or AT effects (evaluating patients according to the treatment they received, not the treatment they were assigned to) may be of specific interest. The conceptual difference of the ITT and PP effects (or estimands) has gained more attention for clinical trial design recently, for example through the ICH E9 (R1) addendum on estimands [4]. With perfect adherence to the treatment strategy and study protocol, ITT, PP, and AT results would be identical. Compared to PP and AT effects, ITT effects are theoretically unbiased estimates of the randomly assigned treatment regardless of the adherence [2], but they may increasingly deviate from PP and AT effects with increasing non-adherence. The reasons for adherence are frequently not random but associated with prognostic factors (e.g. sicker patients may have more difficulties to follow the intended treatment schedule, or they may be more motivated to adhere to the treatment). When there are confounding factors which are associated with both adherence and the outcome of interest, unadjusted PP or AT analyses would be biased. Such confounding factors may be prognostic factors available at baseline, such as age, disease stage, or preferences and values of patients. Standard statistical approaches adjusting for such variables at baseline may, at least theoretically, address some of this confounding. However, there are often also time-varying confounders, which also include the randomly assigned treatments that the study aims to explore [2]. This can, for example, be unsatisfying weight loss leading to non-adherence to a demanding workout intervention. In such cases, standard approaches for confounder control could be inappropriate [5].

Marginal structural model (MSM) analyses are used to adjust for confounding in observational research [5, 6] and they can address time-varying confounding. If the relevant confounders are known, measured and adequately implemented in the modelling [5], MSM should theoretically allow to provide valid estimates of the PP and AT effect.

Beyond conceptual considerations and frameworks, there is to our knowledge no comprehensive empirical evaluation of using MSM analyses in clinical trial research. We conducted a meta-epidemiological analysis aiming to systematically identify situations where MSM analyses have been used in RCTs, understand why these analytical approaches were chosen, how answers to clinical questions agree between these different clinical trial analysis approaches and how this may impact health care decision making [7]. We specifically focused on the relationship of MSM-based effects and results from ITT analyses.

Methods

Search

We conducted four separate searches. First, we searched PubMed using textwords and the medical subject heading for MSM applying the Cochrane sensitivity- and precision-maximizing RCT filter [8] (Webappendix 1). Second, we used the citation search function in Web of Science to screen the titles and abstracts of all articles cited by potentially relevant studies identified through the PubMed search. Third, we screened all references and citations of 12 key references (selected by expert opinion of the authors group) in the field of MSMs [9-20]. Fourth, we also used and updated the search strategy from a related ongoing project in which we compared the effects from non-randomized studies using MSMs with those from systematically identified RCTs not using MSMs. All full-text publications were assessed by two independent reviewers (HE, and one of AL, BS) and disagreements were resolved by discussion or with a third reviewer (LGH).

Selection of studies

We included any RCT (including re-analyses of RCTs) that reported the effects of any health care intervention analyzed using MSM and at least one effect from an ITT, as treated or per protocol analysis. When we were unsure whether a reported effect was analysed using MSM analysis, ITT, or another approach, we asked authors by email for clarification. We contacted authors of 54 trials, in which the use of MSM analysis was not clearly stated but alluded to, to clarify whether or not MSM analysis was used at all and also to analyze the randomized comparison (response rate 52%). For 23 effect estimates from 8 included RCTs [21-28] where we could not clearly determine the ITT effect, we contacted the trial authors for clarification (response rate 88%). We did not verify the methodology of these approaches but relied on the reported description of the methods in the articles or responses to requests, i.e. when the authors described their approach using the words “marginal structural models”, “intention to treat”, “as treated”, “per protocol”, or semantic variations thereof. No other eligibility criteria were applied.

For each eligible RCT, we searched the first publication reporting the results of the primary endpoint (typically the “main” publication). We also searched trial protocols to obtain supplemental information on pre-specification of analyses and to clearly determine the primary outcomes (e.g. by evaluating details of the sample size calculation). To identify these publications, two reviewers (HE, BS) independently screened the reference lists of the MSM-publications, trial homepages, PubMed, and clinicaltrials.gov.

Data extraction

From each eligible RCT, we selected all clearly MSM-based effects on any outcome using any metric (in one case [29], both risk difference and hazard ratio were reported and we extracted only the hazard ratio). For each reported MSM-based effect, we identified any corresponding non-MSM-based effect in the same publication and the main trial publication (where applicable) that was based on the same comparison (i.e. population, intervention, control, outcome) and follow-up time-point (allowing for up to 12 month deviance). We specifically identified any effect from ITT and other analyses such as analyses reported as “per protocol” or “as-treated”. We extracted the MSM-based and corresponding non-MSM-based effect estimates (with 95% confidence intervals), and details on the analysis approaches. For two comparisons of one trial with continuous outcomes, there was no between-group difference and we calculated it using the reported changes from baseline [30, 31]. We extracted the

effects for the overall trial population where possible. In one case, we extracted the results for two mutually exclusive subpopulations (aspirin users and non-users) as no MSM-effect was reported for the overall population [24]. In three other cases, the MSM-based effect was only reported for a subpopulation of the main trial [23, 32, 33] and we only used non-MSM analyses for the same subpopulation.

We extracted general trial characteristics, determined the primary endpoint and whether an MSM analysis was pre-specified according to the protocol or clear statements in the study publications. To determine why MSM analyses are used in RCTs, we extracted any statements on the authors' motivations for using MSMs.

Data analysis

For each eligible trial and outcome, we specifically juxtaposed MSM-based with ITT-based results as well as MSM-based with any other results.

Firstly, using the results from all available analyses, we assessed how frequently treatment effects reported from MSM and other analyses were in the same or in opposite directions, how often there was no overlap between the 95% CIs of the results, and how often the MSM-based effect lay within the 95% CI of the other effects. We also determined the overall vibration of treatment effect estimates per comparison, i.e. the spread between the largest and smallest effect size (on a relative risk [odds ratio or hazard ratio] scale) derived from different analytical methods on the same comparison [34], excluding two trials where we only had effects for continuous outcomes.

Secondly, to specifically focus on MSM-based versus ITT-based results across all comparisons, we selected the main MSM- and main ITT-based effect for each comparison. When multiple variations of such effects were reported, we selected the one described as "main" or "primary" (in the MSM-publication for the MSM-based effect and in the main publication for the ITT effect). When this was unclear, we selected the one first mentioned in the abstract (or in the results section, if none were mentioned in the abstract).

Thirdly, to specifically compare the MSM- and ITT-based results on a trial level, we selected one main comparison of each trial. When there were multiple comparisons on different outcomes in the same trial, we selected the primary outcome or, if unclear, the one first mentioned. For two trials, we selected two comparisons (one trial compared two interventions with one control [35] and another used MSM for two mutually exclusive subpopulations [24]).

We determined if MSM-based relative risk estimates for binary outcomes deviated more or less from the null, i.e. were more or less extreme than ITT-based effects. We tested if one approach more frequently provided more extreme effects than the other with the test for one proportion [36]. We then determined the ratio of these deviations from the null with MSM- versus ITT analysis (by calculating the difference between the deviations on the log-scale and then back transforming to a relative risk scale). For example, when the relative risk estimates are 0.5 and 2.0 with the two approaches, the difference from the null are identical and the ratio of the deviations is 1-fold. A ratio of > 1 indicates more extreme effects for MSM-based results.

Finally, we determined how similar the estimates of MSM- versus ITT analyses are using the ratio of the estimated relative risks (by calculating the absolute difference between MSM-based and ITT-based effect sizes on the log-scale). E.g. if the relative risk estimates are 0.5 and 2.0 with the two approaches,

the ratio of the estimated relative risks is 4-fold. This ratio is >1 by definition as it reflects the absolute difference.

We considered hazard ratios or risk ratios equivalent to odds ratios, when odds ratios were not available. The approximation is sufficiently accurate for modest event rates as those observed in the eligible trials. We used Stata 14.2, R 3.3.2 and Excel 14.0 for all analyses.

Results

The search yielded 4372 records (last searched 19 May 2017), 176 were assessed in full-text. We included 14 publications reporting results of 12 RCTs with a median of 1972 included patients; IQR 870 to 17006) (Figure 1; Table 1). They were published between 2002 and 2016 (median 2013). Six of the 12 RCTs stopped early, 4 for benefit [28, 35, 37, 38] and 2 for harm [27, 39]. The studies evaluated treatment effects of aspirin, anticoagulation, hormone therapy, anticancer drugs, timing of circumcision, antiretrovirals, dietary interventions, antipsychotics, or prevention of mishaps. In 6 of 12 RCTs, the control was inactive, i.e. placebo [27, 28, 39], no intervention [30, 40], or delayed intervention [37]. They reported outcomes related to cardiology [27, 28, 39, 41], oncology [39, 42, 43], infectious diseases [25, 37, 38], diabetes [23], psychiatry [44], gerontology [30], and physical education [40] (Table 1). Double blinding was reported in 6 of 11 RCTs [27, 28, 37-39, 41].

For 7 RCTs [27, 28, 30, 35, 39, 41, 43], we identified a protocol or design paper. The application of MSM was pre-specified only in 1 of the 12 trials [30]. The first or last author of the publication presenting MSM-based results also co-authored the main and, where available, the protocol publication for 9 of 12 trials [38, 39, 43]. The MSM-publication was published a median of 3 years after the main trial publication. The stated motivations for applying MSM were diverse: MSM was used to adjust for "time-dependent" or "time-varying" confounding [22, 24-26, 29, 31-33, 45-47], "non-compliance" or "non-adherence" [24, 25, 29, 31, 40, 45, 47], "loss to follow-up" [22], treatment switching [33], second-line treatment [32], and "to analyze the data as if it were from an observational study rather than a randomized, controlled trial" [23]; Webappendix 2).

Across the 12 RCTs, we identified 24 clinical questions, i.e. comparisons (median 6, IQR 3 to 7 per comparison). Overall, 138 analyses were reported for these 24 comparisons of which 38 were MSM-based (including sensitivity analyses, "crude" and adjusted analyses, different censoring, and different forms of MSM). For 20 of the 24 clinical questions there were ITT analyses reported (in 11 RCTs), AT analyses for 9 (4 RCTs), PP analyses for 3 (1 RCT), and other analyses for 5 (2 RCTs) (Figure 2). Twenty-one comparisons had binary outcomes and 3 comparisons (2 RCTs) had continuous outcomes.

Two analyses using MSM were clearly pre-specified (1 trial), and 4 analyses using MSM were described as "sensitivity analysis" (2 trials). MSM was used to evaluate the primary endpoint in 11 of the 12 RCTs.

Overall relationship of treatment effects

Across all 24 comparisons, the MSM-based results and those from any other reported analyses were all in the same direction in 19 cases (79%), overlapped with all of the 95% CIs (100%), and the MSM-effect lay within all 95% CIs of all other effects in 19 of 22 cases (86%).

Among the 123 analyses reported for 21 comparisons with binary outcomes, the median spread between the largest and smallest effect estimate was 1.19 on a relative risk scale, i.e. the largest effect estimate was 1.19-fold (median; IQR 1.13 to 1.34; Table 2) larger than the smallest.

Relationship of MSM- and ITT-based results

MSM-based and ITT-based results were all in the same direction across all 20 available comparisons (100%; Table 2). Their CIs overlapped in all 18 cases with available CI information (100%), and the MSM-effect lay within the 95% CI of ITT effects in 16 cases (89%). Twelve of 17 (71%, 3 cases with at least 1 CI missing) had both the same direction of effect and were both nominally significant or both nominally non-significant (i.e. both 95% CIs included the null or not; Table 2).

MSM-based effects were more extreme in 13 of 20 comparisons (65%); and in 7 of 20 (35%), ITT-based effects were more extreme (p=0.18). The median deviation from the null of the MSM-based effects was 1.35 (IQR 1.19 to 1.59) and of ITT effects 1.24 (IQR 1.10 to 1.29) on a relative risk scale. On average (median), the ratio of these deviations indicated 1.12-fold more extreme MSM-based effects than the corresponding ITT effects (IQR 0.99 to 1.22; Table 2).

When analyzing only the 13 main comparisons (1 trial had no ITT-based result at all), MSM-based effects were more extreme than ITT-based effects in 7 comparisons (54%). In 46%, ITT-based effects were more extreme (p=0.78). The median deviation from the null of the MSM-based effects was 1.39 (IQR 1.19 to 1.69) and of ITT effects 1.24 (IQR 1.15 to 1.34). Here, MSM-based effects were 1.11-fold more extreme (IQR 0.98 to 1.20; Table 2).

The ratio of the estimated relative risks from MSM and ITT was 1.12-fold (IQR 1.02 to 1.22; Table 2), i.e. half of the MSM-based effects deviated at least 1.12-fold from ITT effects. Among the 11 main comparisons, this was 1.11-fold (IQR 1.02 to 1.20; Table 2).

Discussion

Principal findings

In this empirical analysis, we found 12 trials with 138 effect estimates for 24 clinical questions (comparisons) which reported results from MSM-based and conventional analyses (Figure 2). The main motivations for using MSM were related to missing data and other protocol deviations. The differences between MSM-based and other effects, including ITT effects, were typically within chance and the effects were in the same direction. However, the quantitative differences across reported effects even within the same trial for the same outcome and the same author groups using different methods can be substantial. MSM does not consistently yield more extreme effects than ITT. Overall, MSM and ITT effects were similar, the absolute difference was less than 1.12-fold in half of the comparisons. However, while a difference of 1.12-fold may be modest for some outcomes, it may be clinically very meaningful for others (e.g. death).

The substantial vibration between effect sizes from different analytic methods may be of less relevance for clinical decision-making as all effect estimates had the same direction. However, when quantifying effect estimates and their CIs across several studies (e.g. in meta-analyses, health technology assessments, or indirect comparisons of treatment effects), it may make a substantial difference which analysis method is chosen. When it comes to weighing benefits and harms of treatments or informing

shared decisions, for example when relative risks are translated to numbers needed to treat or harm, variations of effect sizes could matter.

There were on average 6 estimates of the very same outcome (we explicitly searched for trials with at least 2 analyses of the same outcome and many analyses were explorative to demonstrate the analytic approaches). However, when publications offer several effect estimates for one and the same outcome, it may be difficult for healthcare decision makers to know which to base their decisions on. In one study, for example, there were 11 ITT effects, and many studies had two or more ITT effects. The analyses for these estimates followed different approaches and had different degrees of statistical adjustments (e.g. crude and adjusted for various covariates) [48]. Essentially, almost all of the analyses were not clearly pre-specified and this could add uncertainty in the results [48, 49]. In such a setting, the substantial vibration of effects magnifies the impact of selective reporting biases. Post hoc calculations of effect estimates may impact the overall assessment of treatments substantially and further increases the risk for misguided care or policy making. It is also unknown how many additional analyses, with different models and adjustments might also have been performed, yet were not reported at all. Our findings highlight that mere pre-specification of the outcomes (and not specifically the analyses thereof) in clinical trials may not be sufficient to prevent selective reporting bias. Even when the results for an outcome are reported for the same time-point as pre-specified in protocols and trial registries, the results from various statistical approaches may provide different effect sizes and can be selectively reported.

Overall, the spread between the effect estimates from the statistical analyses on the very same outcome was substantial (1.19-fold). In the present sample of trials, the use of MSM was often an explorative approach. However, conducting several analytical methods in addition to the pre-specified analyses, especially methods as complex as MSM that give plenty of options for specification, could increase the risk for selective reporting of only some of the statistical analyses. Even when the approach itself would be pre-specified, statistical details of applying such a complex approach may still have great impact on the results. This complexity requires highly detailed pre-specification because of the possible impact of selective reporting bias, which could possibly be larger than the impact of selecting conceptually different analytical approaches (or estimands [4]).

Comparison with other studies

This is, as far as we know, the first meta-epidemiological analysis comparing the results from causal modelling analyses with conventional analyses within trials across all medical fields. Several empirical studies compared ITT and PP analyses within one medical field or a specific time range [50-52]. A re-analysis of an RCT evaluating interventions for symptom management compared the conclusions from ITT (without imputing missing data) with those from PP analysis [50]. While the conclusions did not differ, the PP analysis also indicated which intervention and dose strategies affected symptoms [50]. A systematic review of RCTs reporting both ITT and PP analyses on a primary binary endpoint found effects from PP analyses more extreme and the ratio to ITT analyses varied greatly (0.39 to 2.53)[51]. In line with our findings, they concluded that protocol deviations can lead to systematic and unpredictable bias and that a trial's conclusion should not be based on the effect of either ITT or PP alone [51]. A meta-epidemiological study compared the results from conventional ITT analyses with those from modified ITT analyses or non-ITT analyses. Similar to our results, they found that the ITT results had less extreme effects [53].

Limitations

Our study has several limitations. First, we only identified 12 trials for which we had MSM- and non-MSM-based effects and that focused on clinical decision making. Many of the excluded RCTs did not use MSM to analyze the randomized comparison but merely used the trial database to evaluate associations of non-randomized exposures or patient characteristics with outcomes.

Second, we encountered various different forms and descriptions of MSM, e.g. standard MSM [16], augmented MSM, adaptively truncated MSM [46], MSM for binary and continuous outcomes [33], using IPTW, IPCW, IPW, G-estimation, adjusted or "unadjusted" [40], censored at different timepoints [25] and adjusted for different covariates with or without two-way interactions between randomization status and each covariate [25]. Some of our included studies even reported the results of multiple different forms of MSM within their study [25, 26, 40, 46].

Third, we did not verify the analytic approaches and relied on the authors' descriptions of them. Although we believe that the authors are probably the best experts for their data and analyses and have correctly classified and described them, details of the definitions may still be inconsistent [54].

Fourth, the trials that applied MSM to analyze the randomized comparison were mainly very large and highly cited (median citation count of main publications 1388 (IQR 142 to 2121; SCOPUS 7 January 2018). All but 3 studies [30, 40, 42] were among the top 1% of the related medical trial literature ("SCOPUS Citation Benchmarking Compared to Medicine articles of same age and document type"; 1 study not found on SCOPUS and not counted [44]). Many (5/12) trials were also discontinued early, more frequently than in the typical clinical trial literature [55]. Hence our sample appears not to be representative of all RCTs.

Fifth, we encountered problems with vague reporting of the analysis methods used. E.g. an analysis was merely described as "conventional Cox model" and it was unclear who was analyzed or how missing patient data were imputed. Several terms are not globally defined, e.g. "patients evaluable for efficacy", "intent-to-treat subset", or "population with observed cases". This may confuse readers as they have different meanings to different people [54].

Sixth, the reporting of adherence, protocol violations, treatment switches, and missing data was typically not sufficiently clear to allow us to consider this in further analyses. We were not able to explore the agreement between effects sizes in relation to these factors.

Seventh, the application of MSM in many studies was for very different reasons. MSM was sometimes applied by highly experienced teams of biostatisticians who developed the approach and who conducted the analyses post-hoc for methodological demonstration purposes and not with the direct intention to inform healthcare decision making. This further adds to the very limited generalizability of this small but nevertheless systematically derived sample of trials.

Finally, for each outcome, we intended to extract information indicating potential problems that may motivate authors to use MSM or other specific models. While we found statements that clearly indicated such issues, the reporting quality was very heterogeneous.

Overall, MSM analyses require more sophisticated modelling than ITT analyses. It is difficult to pre-specify and collect the detailed high-quality data that are required for analyzing all possible post-randomization confounders, such as non-adherence [56]. Since patients' preferences and values

leading to non-adherence are almost never included in data collection, important confounders very likely remain unmeasured and hence cannot be included in the modelling. Therefore, probably there is always some residual confounding bias even in MSM-adjusted effects. Furthermore, caution is required to pre-specify analyses where possible and apply strict safeguards to avoid selective reporting or biases introduced by unblinded analyses. These limitations are less relevant in ITT analyses which don't require such adjustments and are more straightforward to pre-specify. Selective reporting bias may have more impact on results used for decision-making than using conceptually different statistical approaches per se. Without very detailed and strict measures as safeguards to avoid research-associated biases such as selective reporting, the theoretical value of this promising approach may be entirely neutralized under "real world" research conditions.

Conclusion

Overall, we conclude that MSM-based results in randomized trials typically agreed with ITT and other conventional analyses of RCTs. They may theoretically provide very helpful insights and different perspectives in treatment effects, especially when there are high rates of attrition and non-adherence. However, there is a wide spread across all reported effects for the same outcome that requires utmost attention and complex safeguards to prevent selective reporting bias and related problems.

Acknowledgments

We thank the authors of the primary studies for their timely responses to our information requests. We also thank Soheila Aghlmandi (University of Basel) for helping with statistical unclarities, and Benjamin Kasenda (University of Basel) for providing subject matter expertise for oncology topics.

Contributors

HE, LGH conceived the study with input on the study design by JPAI. HE, LGH, BS, AL extracted the data. HE, LGH and JPAI analyzed the data. HE, LGH, JPAI interpreted the results. HE wrote the first draft and all authors made revisions on the manuscript. All authors read and approved the final version of the paper. HE acquired funding for this study. HE and LGH are the guarantors.

Funding

This work was funded by a stipend by the PhD Educational Platform Health Sciences (PPHS). The Meta-Research Innovation Center at Stanford is funded by a grant by the Laura and John Arnold Foundation. The Basel Institute of Clinical Epidemiology and Biostatistics is supported by Stiftung Institut für klinische Epidemiologie.

Role of the funding source

The funders had no role in design and conduct of the study; collection, management, analysis, and interpretation of the data; and preparation, review, or approval of the manuscript or its submission for publication.

Transparency declaration

The Corresponding Author affirms that the manuscript is an honest, accurate, and transparent account of the study being reported; that no important aspects of the study have been omitted; and that any discrepancies from the study as planned have been explained.

Ethical approval

Not required, this article does not contain any personal medical information about any identifiable living individuals.

References

[1] Hollis S, Campbell F. What is meant by intention to treat analysis? Survey of published randomised controlled trials. BMJ (Clinical research ed). 1999;319:670-4.

[2] Hernán MA, Hernández-Díaz S. Beyond the intention to treat in comparative effectiveness research. Clinical trials (London, England). 2012;9:48-55.

[3] Karanicolas PJ, Montori VM, Devereaux PJ, Schünemann H, Guyatt GH. A new 'Mechanistic-Practical' Framework for designing and interpreting randomized trials. Journal of Clinical Epidemiology. 2009;62:479-84.

[4] European Medicines Agency. Draft ICH E9 (R1) addendum on estimands and sensitivity analysis in clinical trials to the guideline on statistical principles for clinical trials, step 2b - Revision 1. 2017.

[5] Williamson T, Ravani P. Marginal structural models in clinical research: when and how to use them? Nephrol Dial Transplant. 2017;32:ii84-ii90.

[6] Delaney JA, Daskalopoulou SS, Suissa S. Traditional versus marginal structural models to estimate the effectiveness of beta-blocker use on mortality after myocardial infarction. Pharmacoepidemiol Drug Saf. 2009;18:1-6.

[7] Hemkens LG, Contopoulos-Ioannidis DG, Ioannidis JPA. Routinely Collected Data and Comparative Effectiveness Evidence: Promises and Limitations. CMAJ (In Press).

[8] Lefebvre C, Manheimer E, J G. Chapter 6: Searching for studies. In: Higgins J, Green S (editors) Cochrane Handbook for Systematic Reviews of Interventions Version 510 (updated March 2011) The Cochrane Collaboration; 2011.

[9] Cole SR, Hernan MA. Constructing inverse probability weights for marginal structural models. American journal of epidemiology. 2008;168:656-64.

[10] Hernan MA, Brumback B, Robins JM. Marginal structural models to estimate the causal effect of zidovudine on the survival of HIV-positive men. Epidemiology (Cambridge, Mass). 2000;11:561-70.

[11] Hernán MA, Brumback B, Robins JM. Marginal Structural Models to Estimate the Joint Causal Effect of Nonrandomized Treatments. Journal of the American Statistical Association. 2001;96:440-8.

[12] Hernan MA, Robins JM. Estimating causal effects from epidemiological data. J Epidemiol Community Health. 2006;60:578-86.

[13] Robins J. Marginal structural models. 1997 Proceedings of the Section on Bayesian Statistical Science. Alexandria: American Statistical Association, 1998;1-10.

[14] Robins JM. Correction for non-compliance in equivalence trials. Stat Med. 1998;17:269-302; discussion 87-9.

[15] Robins JM. Association, causation, and marginal structural models. Synthese. 1999;121:151-79.

[16] Robins JM. Marginal Structural Models versus Structural nested Models as Tools for Causal inference. In: Halloran ME, Berry D, editors. Statistical Models in Epidemiology, the Environment, and Clinical Trials. New York, NY: Springer New York; 2000. p. 95-133.

[17] Robins JM, Greenland S, Hu F-C. Estimation of the Causal Effect of a Time-Varying Exposure on the Marginal Mean of a Repeated Binary Outcome. Journal of the American Statistical Association. 1999;94:687-700.

[18] Robins JM, Hernan MA, Brumback B. Marginal structural models and causal inference in epidemiology. Epidemiology (Cambridge, Mass). 2000;11:550-60.

[19] Suarez D, Borras R, Basagana X. Differences between marginal structural models and conventional models in their exposure effect estimates: a systematic review. Epidemiology (Cambridge, Mass). 2011;22:586-8.

[20] VanderWeele TJ. Marginal structural models for the estimation of direct and indirect effects. Epidemiology (Cambridge, Mass). 2009;20:18-26.

[21] Mehta SD, Moses S, Agot K, Maclean I, Odoyo-June E, Li H, et al. Medical Male Circumcision and Herpes Simplex Virus 2 Acquisition: Posttrial Surveillance in Kisumu, Kenya. Journal of Infectious Diseases. 2013;208:1869-76.

[22] Mehta SD, Moses S, Agot K, Odoyo-June E, Li H, Maclean I, et al. The long-term efficacy of medical male circumcision against HIV acquisition. Aids. 2013;27:2899-907.

[23] Salas-Salvado J, Bullo M, Estruch R, Ros E, Covas M-I, Ibarrola-Jurado N, et al. Prevention of Diabetes With Mediterranean Diets A Subgroup Analysis of a Randomized Trial. Annals of Internal Medicine. 2014;160:1-+.

[24] Alexander JH, Lopes RD, Thomas L, Alings M, Atar D, Aylward P, et al. Apixaban vs. warfarin with concomitant aspirin in patients with atrial fibrillation: insights from the ARISTOTLE trial. Eur Heart J. 2014;35:224-32.

[25] Cain LE, Cole SR. Inverse probability-of-censoring weights for the correction of time-varying noncompliance in the effect of randomized highly active antiretroviral therapy on incident AIDS or death. Stat Med. 2009;28:1725-38.

[26] Cook NR, Cole SR, Hennekens CH. Use of a marginal structural model to determine the effect of aspirin on cardiovascular mortality in the Physicians' Health Study. Am J Epidemiol. 2002;155:1045-53.

[27] Ridker PM, Cook NR, Lee IM, Gordon D, Gaziano JM, Manson JE, et al. A randomized trial of low-dose aspirin in the primary prevention of cardiovascular disease in women. N Engl J Med. 2005;352:1293-304.

[28] STEERING COMMITTEE of THE PHYSICIANS' HEALTH STUDY RESEARCH GROUP. Final report on the aspirin component of the ongoing Physicians' Health Study. N Engl J Med. 1989;321:129-35.

[29] Toh S, Hernandez-Diaz S, Logan R, Rossouw JE, Hernan MA. Coronary heart disease in postmenopausal recipients of estrogen plus progestin therapy: does the increased risk ever disappear? A randomized trial. Ann Intern Med. 2010;152:211-7.

[30] Ravussin E, Redman LM, Rochon J, Das SK, Fontana L, Kraus WE, et al. A 2-Year Randomized Controlled Trial of Human Caloric Restriction: Feasibility and Effects on Predictors of Health Span and Longevity. J Gerontol A Biol Sci Med Sci. 2015;70:1097-104.

[31] Rochon J, Bhapkar M, Pieper CF, Kraus WE. Application of the Marginal Structural Model to Account for Suboptimal Adherence in a Randomized Controlled Trial. Contemp Clin Trials Commun. 2016;4:222-8.

[32] Yamaguchi T, Ohashi Y. Adjusting for differential proportions of second-line treatment in cancer clinical trials. Part II: an application in a clinical trial of unresectable non-small-cell lung cancer. Stat Med. 2004;23:2005-22.

[33] Faries D, Ascher-Svanum H, Belger M. Analysis of treatment effectiveness in longitudinal observational data. J Biopharm Stat. 2007;17:809-26.

[34] Ioannidis JP. Why most discovered true associations are inflated. Epidemiology. 2008;19:640-8.

[35] Estruch R, Ros E, Salas-Salvado J, Covas MI, Corella D, Aros F, et al. Primary prevention of cardiovascular disease with a Mediterranean diet. N Engl J Med. 2013;368:1279-90.

[36] Social Science Computing Cooperative. Stata for Students: Proportion Tests. 2016.

[37] Bailey RC, Moses S, Parker CB, Agot K, Maclean I, Krieger JN, et al. Male circumcision for HIV prevention in young men in Kisumu, Kenya: a randomised controlled trial. Lancet. 2007;369:643-56.

[38] Hammer SM, Squires KE, Hughes MD, Grimes JM, Demeter LM, Currier JS, et al. A controlled trial of two nucleoside analogues plus indinavir in persons with human immunodeficiency virus infection and CD4 cell counts of 200 per cubic millimeter or less. AIDS Clinical Trials Group 320 Study Team. N Engl J Med. 1997;337:725-33.

[39] Rossouw JE, Anderson GL, Prentice RL, LaCroix AZ, Kooperberg C, Stefanick ML, et al. Risks and benefits of estrogen plus progestin in healthy postmenopausal women: principal results From the Women's Health Initiative randomized controlled trial. JAMA. 2002;288:321-33.

[40] Ranapurwala SI, Denoble PJ, Poole C, Kucera KL, Marshall SW, Wing S. The effect of using a pre-dive checklist on the incidence of diving mishaps in recreational scuba diving: a cluster-randomized trial. Int J Epidemiol. 2015.

[41] Granger CB, Alexander JH, McMurray JJ, Lopes RD, Hylek EM, Hanna M, et al. Apixaban versus warfarin in patients with atrial fibrillation. N Engl J Med. 2011;365:981-92.

[42] Negoro S, Masuda N, Takada Y, Sugiura T, Kudoh S, Katakami N, et al. Randomised phase III trial of irinotecan combined with cisplatin for advanced non-small-cell lung cancer. Br J Cancer. 2003;88:335-41.

[43] Patel JD, Socinski MA, Garon EB, Reynolds CH, Spigel DR, Olsen MR, et al. PointBreak: a randomized phase III study of pemetrexed plus carboplatin and bevacizumab followed by maintenance pemetrexed and bevacizumab versus paclitaxel plus carboplatin and bevacizumab followed by maintenance bevacizumab in patients with stage IIIB or IV nonsquamous non-small-cell lung cancer. J Clin Oncol. 2013;31:4349-57.

[44] Tunis SL, Faries DE, Nyhuis AW, Kinon BJ, Ascher-Svanum H, Aquila R. Cost-effectiveness of olanzapine as first-line treatment for schizophrenia: Results from a randomized, open-label, I-year trial. Value in Health. 2006;9:77-89.

[45] Cook NR, Cole SR, Buring JE. Aspirin in the primary prevention of cardiovascular disease in the Women's Health Study: effect of noncompliance. Eur J Epidemiol. 2012;27:431-8.

[46] Bai X, Liu J, Li L, Faries D. Adaptive truncated weighting for improving marginal structural model estimation of treatment effects informally censored by subsequent therapy. Pharm Stat. 2015;14:448-54.

[47] Toh S, Hernandez-Diaz S, Logan R, Robins JM, Hernan MA. Estimating absolute risks in the presence of nonadherence: an application to a follow-up study with baseline randomization. Epidemiology. 2010;21:528-39.

[48] Saquib N, Saquib J, Ioannidis JP. Practices and impact of primary outcome adjustment in randomized controlled trials: meta-epidemiologic study. BMJ. 2013;347:f4313.

[49] Ioannidis JP, Caplan AL, Dal-Re R. Outcome reporting bias in clinical trials: why monitoring matters. BMJ. 2017;356:j408.

[50] Given B, Given CW, Sikorskii A, You M, McCorkle R, Champion V. Analyzing symptom management trials: the value of both intention-to-treat and per-protocol approaches. Oncol Nurs Forum. 2009;36:E293-302.

[51] Porta N, Bonet C, Cobo E. Discordance between reported intention-to-treat and per protocol analyses. J Clin Epidemiol. 2007;60:663-9.

[52] Schiffner R, Schiffner-Rohe J, Gerstenhauer M, Hofstadter F, Landthaler M, Stolz W. Differences in efficacy between intention-to-treat and per-protocol analyses for patients with psoriasis vulgaris and atopic dermatitis: clinical and pharmacoeconomic implications. Br J Dermatol. 2001;144:1154-60.

[53] Abraha I, Cherubini A, Cozzolino F, De Florio R, Luchetta ML, Rimland JM, et al. Deviation from intention to treat analysis in randomised trials and treatment effect estimates: meta-epidemiological study. BMJ. 2015;350:h2445.

[54] Alshurafa M, Briel M, Akl EA, Haines T, Moayyedi P, Gentles SJ, et al. Inconsistent definitions for intention-to-treat in relation to missing outcome data: systematic review of the methods literature. PLoS One. 2012;7:e49163.

[55] 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;303:1180-7.

[56] Hernan MA, Robins JM. Per-Protocol Analyses of Pragmatic Trials. N Engl J Med. 2017;377:1391-8.

[57] Lopes RD, Alexander JH, Al-Khatib SM, Ansell J, Diaz R, Easton JD, et al. Apixaban for reduction in stroke and other ThromboembоLic events in atrial fibrillation (ARISTOTLE) trial: design and rationale. Am Heart J. 2010;159:331-9.

[58] Hennekens CH, Eberlein K. A randomized trial of aspirin and beta-carotene among U.S. physicians. Prev Med. 1985;14:165-8.

[59] Patel JD, Bonomi P, Socinski MA, Govindan R, Hong S, Obasaju C, et al. Treatment rationale and study design for the pointbreak study: a randomized, open-label phase III study of pemetrexed/carboplatin/bevacizumab followed by maintenance pemetrexed/bevacizumab versus paclitaxel/carboplatin/bevacizumab followed by maintenance bevacizumab in patients with stage IIIB or IV nonsquamous non-small-cell lung cancer. Clin Lung Cancer. 2009;10:252-6.

[60] The Women's Health Initiative Study Group. Design of the Women's Health Initiative clinical trial and observational study. The Women's Health Initiative Study Group. Control Clin Trials. 1998;19:61-109.

[61] Buring JE, Hennekens CH. The Women's Health Study: Summary of the study design. J Myocardial Ischemia. 1992;4:27-9.

Tables

Table 1: Characteristics of included studies

RCT	No. Randomized	Patients' Condition	Intervention and control	Total number of pertinent comparisons	Outcomes with MSM-based results Analytic approach (n)
ACTG 320* [25, 38]	1156	HIV positive, immunosuppressed, ART-experienced patients	HAART (Zidovudine and Lamivudine plus Indinavir) vs CART (Zidovudine and Lamivudine)	11	AIDS or death (primary) MSM (8) ITT (3)
ARISTOTLE [24, 41, 57]	18201 (using aspirin at BL: 5632)	Atrial fibrillation (in aspirin users and non-users)	Apixaban vs Warfarin	6	Stroke or systemic embolism (primary, subgroups only) MSM (1) ITT (2) Major bleeding MSM (1) As treated (2)
	18201 (Not using aspirin at BL: 12569			6	Stroke or systemic embolism (primary, subgroups only) MSM (1) ITT (2) Major bleeding MSM (1) As treated (2)
CALERIE [30, 31]	220	Healthy, young- and middle-aged nonobese men and women	Calorie restriction (behavioral approach with dietary modifications) vs no calorie goal (no dietary or behavioral counseling)	6	RMR (primary) MSM (1) ITT (2) Core temperature (primary) MSM (1) ITT (2)

Kisumu * [21, 22, 37]	2784	Uncircumcised, HIV-negative young men	Immediate vs delayed circumcision	6	HIV incidence (primary) MSM (1) As treated (2) Herpes simplex virus 2 incidence MSM (1) As treated (2)
Negoro / Yamaguchi [32, 42] ***	398 (MSM analysis only for 2 of 3 groups with 266 patients)	Stage IIIB lung cancer (NSCLC)	Irinotecan hydrochloride vs cisplatin	4	Overall survival (primary) MSM (1) ITT (3)
PHS * [26, 28, 58]	22071	Male physicians	Aspirin vs placebo	10	Cardiovascular mortality (primary) MSM (3) ITT (4) As treated (3)
PointBreak [43, 46, 59]	939	Stage IIIB or IV lung cancer (NSCLC)	"Pemetrexed/Carboplatin/Bevacizumab followed by maintenance Pemetrexed/ Bevacizumab" vs "Paclitaxel/Carboplatin/Bevacizumab Followed by Maintenance Bevacizumab"	6	Overall survival (primary) MSM (3) ITT (3)
PREDIMED * [23, 35]	7447 (Non-diabetic subgroup: 3833)	Risk factors for CVD	Mediterranean diet supplemented with extra-virgin olive oil vs advice on a low-fat diet	12	Type 2 diabetes mellitus incidence MSM (1) ITT (11)
			Mediterranean diet supplemented with nuts vs advice on a low-fat diet	12	Type 2 diabetes mellitus incidence MSM (1) ITT (11)
Ranapurwala [40]	1660	Recreational scuba divers	Checklist vs no checklist	18	Any diving mishap (primary)

	(70 randomized units)				MSM (2) ITT (2) Per protocol (2) Major diving mishaps MSM (2) ITT (2) Per protocol (2) Minor diving mishaps MSM (2) ITT (2) Per protocol (2)
Tunis / Faries [33, 44]	664 (MSM analysis only for 2 of 3 groups with 443 patients)	Patients with schizophrenia or schizoaffective disorder	Olanzapine vs “fail-first” algorithm on conventional	9	Change in brief psychiatric rating scale (primary) MSM (1) ITT (2) On drug (4) Epoch (2)
WHI * [29, 39, 60]	16608	Postmenopausal women with intact uterus	Estrogen-plus-progestin vs placebo	7	Coronary Heart Disease (primary) MSM (1) ITT (2) Invasive breast cancer incidence MSM (1) ITT (3)
WHS * [27, 45, 61]	39876	Female health professionals	Aspirin vs placebo	25	Major cardiovascular events (including myocardial infarction, stroke, cardiovascular disease mortality) (primary) MSM (1) ITT (4) As treated (2) On drug (1) Myocardial infarction MSM (1) ITT (3)

	As treated (1) On drug (1) Stroke MSM (1) ITT (3) As treated (1) On drug (1) Cardiovascular disease mortality MSM (1) ITT (3) As treated (1)

* Trial stopped early

** The main publication is based on 2 year follow up (effects not considered)

*** The original study population were patients with untreated NSCLC stage IIIB and IV, however, MSM-based results are only available for stage III patients. Also, the original comparison of the study consisted of 3 treatment arms of which 2 were different doses of Irinotecan. As MSM-based results were only available for the comparison with the lower dose Irinotecan (60 mg m^{-2}), we do not present the third arm (100 mg m^{-2}).

ACTG 320: AIDS Clinical Trial Group; AIDS: Acquired Immune Deficiency Syndrome; ARISTOTLE: Apixaban for Reduction in Stroke and Other Thromboembolic Events in Atrial Fibrillation trial; CALERIE: Comprehensive Assessment of the Long-term Effects of Reducing Intake of Energy; HIV: human immunodeficiency virus; MSM: Marginal structural models; PHS: Physicians' Health Study; PREDIMED: Primary Prevention of Cardiovascular Disease with a Mediterranean Diet; WHI: Women's Health Initiative; WHS: Women's Health Study

Table 2 Relationship of effect estimates per outcome

Comparison	Vibration of treatment effect estimates: Spread of lowest vs. highest relative risk estimate across all reported analyses*	Ratio of deviations from the null with MSM and ITT (x-fold more extreme effects with MSM)*#	Ratio of the relative risks with MSM and ITT* (x-fold difference between effects)	MSM and ITT effects in same direction	MSM and ITT effects with same stat. significance	MSM and ITT effect CI overlapping	MSM effect within CI of ITT effect
ACTG 320: AIDS or death	1,13	1,11	1,11	yes	yes	yes	yes
ARISTOTLE (Aspirin non-users): Major Bleeding	1,05	NA	NA	NA (no ITT effect)	NA (no ITT effect)	NA (no ITT effect)	NA (no ITT effect)
ARISTOTLE (Aspirin users): Major Bleeding	1,04	NA	NA	NA (no ITT effect)	NA (no ITT effect)	NA (no ITT effect)	NA (no ITT effect)
ARISTOTLE (Aspirin non-users): Stroke or systemic embolism	1,02	0,98	1,02	yes	yes	yes	yes
ARISTOTLE (Aspirin users): Stroke or systemic embolism	1,24	1,22	1,22	yes	yes	yes	yes
Kisumu: HIV incidence	1,18	NA	NA	NA (no ITT effect)	NA (no ITT effect)	NA (no ITT effect)	NA (no ITT effect)
Kisumu: HSV-2 incidence	1,14	NA	NA	NA (no ITT effect)	NA (no ITT effect)	NA (no ITT effect)	NA (no ITT effect)
Negoro/Yamaguchi: Survival	1,03	0,97	1,03	yes	yes	yes	yes
PREDIMED (EVOO): Incidence type 2 diabetes mellitus	2,09	0,99	1,01	yes	yes	yes	yes
PREDIMED (Nuts): Incidence type 2 diabetes mellitus	1,86	0,99	1,01	yes	yes	yes	yes
PHS: CVD mortality	1,42	1,3	1,3	yes	yes	yes	yes
PointBreak: Overall survival	2,06	1,12	1,12	yes	NA**	NA**	yes
Ranapurwala: All mishaps	1,18	1,18	1,18	yes	no	yes	yes
Ranapurwala: Major mishaps	1,19	1,16	1,16	yes	no	yes	yes
Ranapurwala: Minor mishaps	1,19	1,19	1,19	yes	yes	yes	yes
WHI: Coronary Heart Disease	1,37	1,31	1,31	yes	no	yes	no

WHI: Invasive breast cancer	1,34	1,33	1,33	yes	yes	yes	no
WHS: CVD Mortality	1,24	1,24	1,24	yes	yes	yes	yes
WHS: Major CVD events	1,15	0,98	1,02	yes	yes	yes	yes
WHS: Myocardial infarction	1,09	1,09	1,09	yes	yes	yes	yes
WHS: Stroke	1,19	0,98	1,02	yes	no	yes	yes
Tunis/Faries: change in BPRS	NA	NA	NA	yes	no	yes	yes
CALERIE: resting metabolic rate	NA	NA	NA	yes	NA***	NA***	NA***
CALERIE: Core Temperature	NA	NA	NA	yes	NA***	NA***	NA***
Median (IQR) or Total (%)	1.19 (1.13 to 1.34) ****	1.12 (0.99 to 1.22)	1.12 (1.02 to 1.22)	Yes: 20/20 (100 %) No: 0/24 (0 %)	Yes: 12/17 (71 %) No: 5/17 (29 %)	Yes: 17/17 (100%) No: 0/17 (0%)	Yes: 16/18 (89%) No: 2/18 (11%)
Median (IQR) or Total (%) (main outcomes only)	1.21 (1.15 to 1.53)	1.11 (0.98 to 1.20)	1.11 (1.02 to 1.20)	Yes: 13/13 (100 %) No: 0/13 (0 %)	Yes: 8/11 (73 %) No: 3/11 (27 %)	Yes: 11/11 (100%) No: 0/11 (0%)	Yes: 11/12 (92%) No: 1/12 (8%)

*) dichotomous outcomes only

**) No 95% confidence interval for the MSM-based result reported

***) No 95% confidence interval for the MSM-based nor the ITT-based result reported

****) The median (IQR) excluding sensitivity analyses is 1.17 (1.08 to 1.24)

#) > 1 indicates more extreme effects for MSM-based results

ACTG 320: AIDS Clinical Trial Group; AIDS: Acquired Immune Deficiency Syndrome; ARISTOTLE: Apixaban for Reduction in Stroke and Other Thromboembolic Events in Atrial Fibrillation trial; CALERIE: Comprehensive Assessment of the Long-term Effects of Reducing Intake of Energy; CI: confidence interval; HIV: human immunodeficiency virus; CVD: cardiovascular disease; EVOO: extra virgin olive oil; IQR: interquartile range; ITT: intention-to-treat; MSM: marginal structural models; NA: not applicable; PHS: Physicians' Health Study; PREDIMED: Primary Prevention of Cardiovascular Disease with a Mediterranean Diet; WHI: Women's Health Initiative; WHS: Women's Health Study

Figures

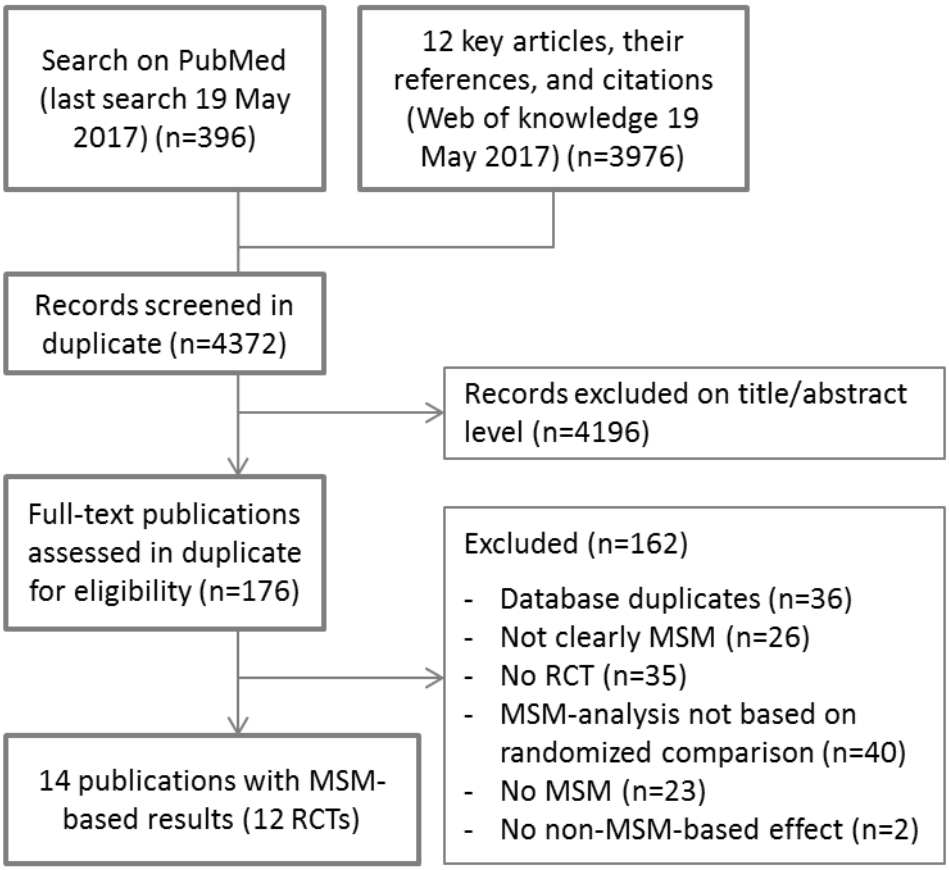

Figure 1: Study flow

MSM: marginal structural models; RCT: randomized controlled trial

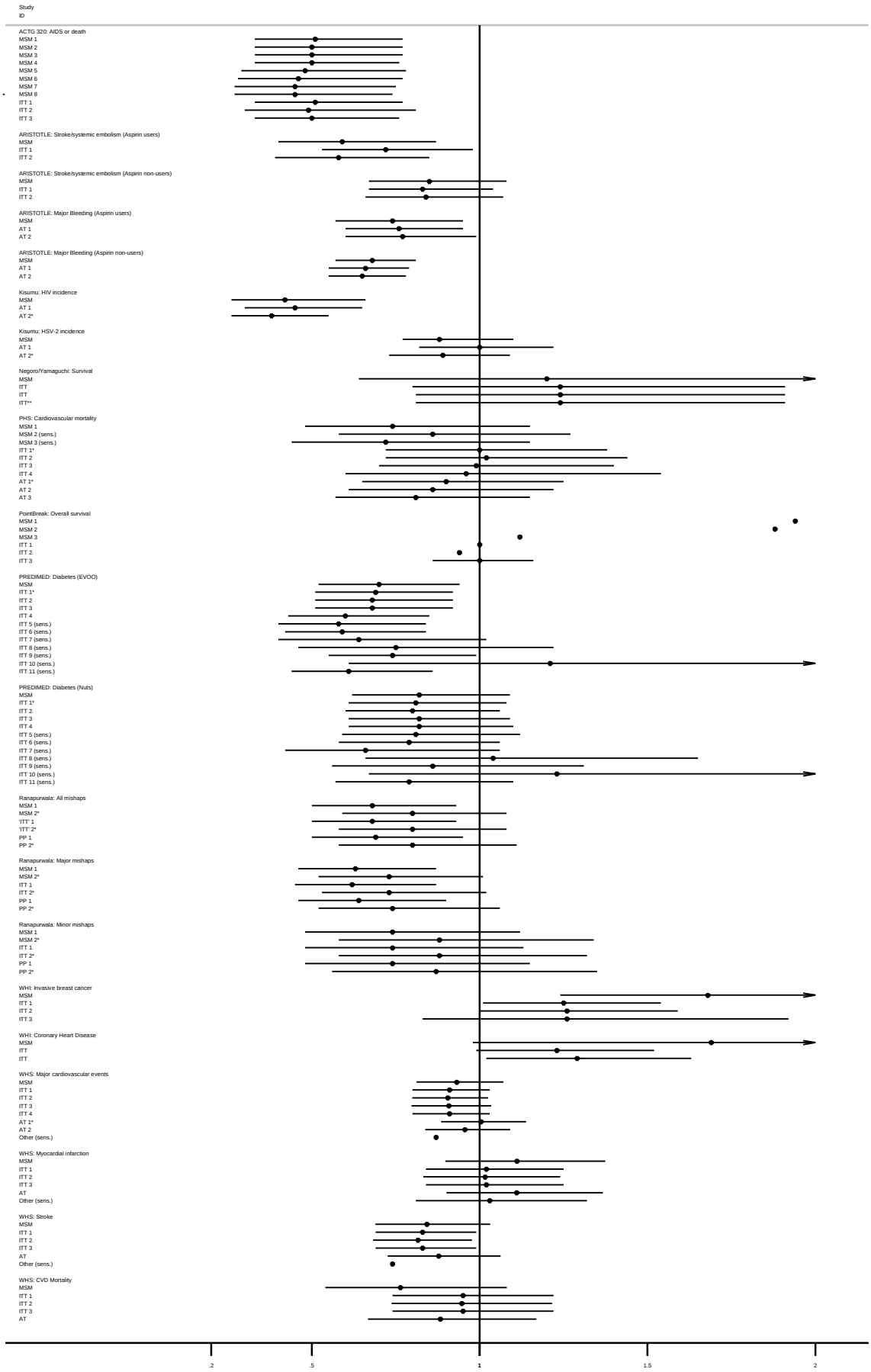
Study
ID
ACTG 320: AIDS or death
MSM 1
MSM 2
MSM 3
MSM 4
MSM 5
MSM 6
MSM 7
MSM 8
ITT 1
ITT 2
ITT 3
ARISTOTLE: Stroke/systemic embolism (Aspirin users)
MSM
ITT 1
ITT 2
ARISTOTLE: Stroke/systemic embolism (Aspirin non-users)
MSM
ITT 1
ITT 2
ARISTOTLE: Major Bleeding (Aspirin users)
MSM
AT 1
AT 2
ARISTOTLE: Major Bleeding (Aspirin non-users)
MSM
AT 1
AT 2
Kisumu: HIV incidence
MSM
AT 1
AT 2*
Kisumu: HSV-2 incidence
MSM
AT 1
AT 2*
Negoro/Yamaguchi: Survival
MSM
ITT
ITT
ITT**
PHS: Cardiovascular mortality
MSM 1
MSM 2 (sens.)
MSM 3 (sens.)
ITT 1*
ITT 2
ITT 3
ITT 4
AT 1*
AT 2
AT 3
PointBreak: Overall survival
MSM 1
MSM 2
MSM 3
ITT 1
ITT 2
ITT 3
PREDIMED: Diabetes (EVOO)
MSM
ITT 1*
ITT 2
ITT 3
ITT 4
ITT 5 (sens.)
ITT 6 (sens.)
ITT 7 (sens.)
ITT 8 (sens.)
ITT 9 (sens.)
ITT 10 (sens.)
ITT 11 (sens.)
PREDIMED: Diabetes (Nuts)
MSM
ITT 1*
ITT 2
ITT 3
ITT 4
ITT 5 (sens.)
ITT 6 (sens.)
ITT 7 (sens.)
ITT 8 (sens.)
ITT 9 (sens.)
ITT 10 (sens.)
ITT 11 (sens.)
Ranapurwala: All mishaps
MSM 1
MSM 2*
'ITT' 1
'ITT' 2*
PP 1
PP 2*
Ranapurwala: Major mishaps
MSM 1
MSM 2*
ITT 1
ITT 2*
PP 1
PP 2*
Ranapurwala: Minor mishaps
MSM 1
MSM 2*
ITT 1
ITT 2*
PP 1
PP 2*
WHI: Invasive breast cancer
MSM
ITT 1
ITT 2
ITT 3
WHI: Coronary Heart Disease
MSM
ITT
ITT
WHS: Major cardiovascular events
MSM
ITT 1
ITT 2
ITT 3
ITT 4
AT 1*
AT 2
Other (sens.)
WHS: Myocardial infarction
MSM
ITT 1
ITT 2
ITT 3
AT
Other (sens.)
WHS: Stroke
MSM
ITT 1
ITT 2
ITT 3
AT
Other (sens.)
WHS: CVD Mortality
MSM
ITT 1
ITT 2
ITT 3
AT
.2
.5
1
1.5
2

Figure 2: Effect estimates and 95% confidence intervals on a relative risk scale reported in the main publication and the publication with MSM-results on the same clinical question (population, intervention, control, outcome, timepoint)

* unadjusted analyses

ACTG 320: AIDS Clinical Trial Group; AIDS: Acquired Immune Deficiency Syndrome; ARISTOTLE: Apixaban for Reduction in Stroke and Other Thromboembolic Events in Atrial Fibrillation trial; AS: as treated; CALERIE: Comprehensive Assessment of the Long-term Effects of Reducing Intake of Energy; CVD: cardiovascular disease; HIV: human immunodeficiency virus; EVOO: extra virgin olive oil; ITT: intention-to-treat; MSM: Marginal structural models; PHS: Physicians' Health Study; PP: per protocol; PREDIMED: Primary Prevention of Cardiovascular Disease with a Mediterranean Diet; Sens.: Sensitivity analysis; WHI: Women's Health Initiative; WHS: Women's Health Study

Webappendix 1

Search details

Key Articles on marginal structural models
1. Cole, Stephen R.; Hernan, Miguel A. Constructing inverse probability weights for marginal structural models. AMERICAN JOURNAL OF EPIDEMIOLOGY Volume: 168 Issue: 6 Pages: 656-664 Published: SEP 15 2008
2. Hernan, MA; Brumback, B; Robins, JM. Marginal structural models to estimate the causal effect of zidovudine on the survival of HIV-positive men. EPIDEMIOLOGY Volume: 11 Issue: 5 Pages: 561-570 DOI: 10.1097/00001648-200009000-00012 Published: SEP 2000
3. Hernan, MA; Brumback, B; Robins, JM. Marginal structural models to estimate the joint causal effect of nonrandomized treatments. JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION Volume: 96 Issue: 454 Pages: 440-448 DOI: 10.1198/016214501753168154 Published: JUN 2001
4. Hernan, MA; Robins, JM Estimating causal effects from epidemiological data. JOURNAL OF EPIDEMIOLOGY AND COMMUNITY HEALTH Volume: 60 Issue: 7 Pages: 578-586 DOI: 10.1136/jech.2004.029496 Published: JUL 2006
5. Robins JM, Hernán MA, Brumback B. Marginal structural models and causal inference in epidemiology. Epidemiology. 2000 Sep;11(5):550-60. PubMed PMID:10955408.
6. Robins JM. Correction for non-compliance in equivalence trials. Stat Med 1998;17:269–302.
7. Robins JM. Marginal structural models versus structural nested models as tools for causal inference. In: Halloran E, Berry D, eds. Statistical Models in Epidemiology: The Environment and Clinical Trials. New York: Springer-Verlag, 1999;95–134.
8. Robins JM. Marginal structural models. In: 1997 Proceedings of the Section on Bayesian Statistical Science, Alexandria, VA: American Statistical Association, 1998;1–10.
9. Robins, JM Association, causation, and marginal structural models. SYNTHESE Volume: 121 Issue: 1-2 Pages: 151-179 DOI: 10.1023/A:1005285815569 Published: NOV 1999
10. Robins, JM; Greenland, S; Hu, FC. Estimation of the causal effect of a time-varying exposure on the marginal mean of a repeated binary outcome. JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION Volume: 94 Issue: 447 Pages: 687-700 DOI: 10.2307/2669978 Published: SEP 1999
11. VanderWeele, Tyler J. Marginal Structural Models for the Estimation of Direct and Indirect Effects. EPIDEMIOLOGY Volume: 20 Issue: 1 Pages: 18-26 Published: JAN 2009
12. Suarez D, Borras R, Basagana X. Differences between marginal structural models and conventional models in their exposure effect estimates: a systematic review. Epidemiology (Cambridge, Mass.). 2011;22(4):586-588.
Search on Pubmed, 2 June 2016
(IPTW[tiab] OR "inverse probability"[tiab] OR (marginal[tiab] AND structur*[tiab] AND model*[tiab]) OR (marginal[tiab] AND "models, structural"[MeSH Terms])) NOT (ANIMALS[MH] NOT HUMANS[MH]) AND (randomized controlled trial[pt] OR controlled clinical trial[pt] OR randomized[tiab] OR placebo[tiab] OR "clinical trials as topic"[MeSH Terms:noexp] OR randomly[tiab] OR trial[ti] NOT ("animals"[MeSH Terms] NOT "humans"[MeSH Terms]))

Webappendix 2

Included RCT	MSM used for	Statements
ACTG 320 [25]	Time-dependent confounding Non-compliance	“Inverse probability-of-censoring weights for the correction of **time-varying noncompliance** in the effect of randomized highly active antiretroviral therapy on incident AIDS or death”
ARISTOTLE [24]	Time-dependent confounding Non-compliance	"In addition to adjusting for baseline variables associated with either the propensity to use aspirin or the outcomes of interest, we used marginal structural models to further **adjust for potential time-dependent confounders** to estimate a HR for the effect of aspirin on outcome, and to test for interaction between aspirin use and the randomized treatment (apixaban vs. warfarin)." “Results from marginal structural model analyses that **accounted for whether a patient was actually taking aspirin** at the time of their bleeding event resulted in similar findings”
CALERIE [31]	Time-dependent confounding Non-adherence	“Application of the marginal structural model to account for suboptimal **adherence** in a randomized controlled trial” “First, stepwise linear regression was used to model the observed percent weight loss, while stepwise logistic regression model was applied to model early **discontinuation** from the intervention.” “This model is complicated and requires careful attention to detail. Which variables to force into the ancillary models, how to construct interaction terms, and how to address **time-dependent covariates** must be considered.” “However, **adherence** is an endogenous variable and satisfies the definition of a **time-dependent confounder**.” “The dataset is then analyzed using a logistic regression model including a term for the time interval as well as the fixed and **time-dependent covariates** of interest.”
Kisumu [21, 22]	Time-dependent confounding and loss to follow-up	"Marginal structural models reduce the bias introduced by self-selection to become circumcised through application of stabilized weighting at each time point for the time-dependent confounders." "For our marginal structural approach, we generated the above described stabilized IPTW and stabilized inverse-probability-of-censoring-weights (IPCWs) **to account for time-dependent confounding and loss to follow-up**."
Negoro/Yamaguchi [32]	Time-dependent confounding Secondline treatment	“In our companion paper [8], we propose to use structural nested models (SNMs) [9, 10] and marginal structural models (MSMs) [10–12] to adjust for **differential proportions of second-line treatment**. In this paper, we deal with clinical application of the two models in detail.”

		"Unlike the usual time-dependent Cox model, the marginal structural Cox model can be used to obtain valid causal inference for the effect of time-varying treatment in the presence of **time-dependent confounders** which satisfy the condition (i) and (ii) introduced in Section 3.1."
PHS [26]	Time-dependent confounding	"The authors used a marginal structural model with **time-dependent inverse probability weights** to estimate the underlying causal effect of aspirin on cardiovascular mortality." "For comparison to the estimates derived from the marginal structural models, we also estimated the effects of aspirin from standard intention-to-treat and as-treated analyses. For comparability, we used pooled logistic regression in these analyses, both with and without the usual **adjustment for time-varying covariates** in time-dependent models."
PointBreak [46]	Time-dependent confounding	"Marginal structural models (MSMs) have been applied to estimate causal treatment effects even in the presence of **time-dependent confounders**."
PREDIMED [23]	To analyze the data As if it were from an observational study rather than a randomized, controlled trial	"We used the marginal structural model to provide the results of an alternative technique to analyze the data as if it were from an observational study rather than a randomized, controlled trial."
Ranapurwala [40]	Non-adherence	"Marginal structural models were used **to account for non-adherence**."
Tunis/Faries [33]	Treatment switching Time-dependent confounding	"Various methods of **eliminating the switching**, such as epoch analyses and on-drug subset analyses, along with use of marginal structural models generated reasonably consistent non-zero treatment effect estimates." "The MSM retains the repeated measures structure of the data and directly **addresses time-varying covariates** while the epoch approaches handles such variables as baseline confounders for the next episode and assesses a different parameter (change from baseline to endpoint of a naturalistic episode of treatment)." "We were particularly interested in the performance of marginal structural modeling (MSM)—as this approach utilizes all of the study data and produces consistent estimates of the causal effect of treatments, even when there are **treatment switching** and **time-varying confounders**."
WHI [29, 47]	Time-dependent confounding Non-adherence	"Inverse probability weighting of marginal structural models has been used **to adjust for nonadherence**, but most studies have provided only relative measures of risk."[47] "Therefore there is no need to estimate separate inverse probability weights to adjust for selection bias due to artificial censoring because the treatment weights estimated in the primary analysis already adjust for the potential **time-varying selection bias** due to artificial censoring."[47] "**Adherence-adjusted** hazard ratios and CHD-free survival curves estimated through inverse probability weighting."[29]

WHS [45]	Time-dependent confounding Non-compliance	"We used marginal structural models (MSMs) to estimate the etiologic effect of continuous aspirin use on CVD events among 39,876 apparently healthy female health professionals aged 45 years and older in the Women's Health Study, a randomized trial of 100 mg aspirin every other day versus placebo." "MSMs, which **adjusted for non-compliance**, were similar for total CVD (HR = 0.93; 95 % CI: 0.81, 1.07) but suggested lower CVD mortality with aspirin use (HR = 0.76; 95 % CI: 0.54, 1.08)." "Marginal structural models (MSMs) [9] can be used to effectively adjust **for time-varying confounding** by nonfatal CVD events which are also affected by aspirin use." "MSMs were used to estimate the etiologic effect of aspirin in the presence of **time-dependent confounders** that are themselves affected by previous aspirin use"

ACTG 320: AIDS Clinical Trial Group; AIDS: Acquired Immune Deficiency Syndrome; ARISTOTLE: Apixaban for Reduction in Stroke and Other Thromboembolic Events in Atrial Fibrillation trial; CALERIE: Comprehensive Assessment of the Long-term Effects of Reducing Intake of Energy; MSM: marginal structural models; PHS: Physicians' Health Study; PREDIMED: Primary Prevention of Cardiovascular Disease with a Mediterranean Diet; WHI: Women's Health Initiative; WHS: Women's Health Study

Webappendix 3

Details of study results from main MSM-based and ITT-based analyses for main comparisons of the 12 trials

ARISTOTLE

The ARISTOTLE trial investigated the effect of apixaban versus warfarin in aspirin users and non-users on stroke or systemic embolism (Table 1). The main MSM-based result was a hazard ratio (HR) of 0.59 (95% CI 0.4 to 0.87) in aspirin users and a HR of 0.85 (95% CI 0.67 to 1.08) in aspirin non-users [24]. We identified an ITT analysis with a HR of 0.72 (95% CI 0.53 to 0.98) for aspirin users and a HR of 0.83 (95% CI 0.67 to 1.04) in aspirin non-users [41].

Physicians' Health Study (PHS)

The Physicians' Health Study (PHS) explored the effects of aspirin compared to placebo on cardiovascular mortality (Table 1). The main MSM-based result was a relative risk (RR) of 0.74 (95% CI 0.48 to 1.15) favoring aspirin [26]. The main ITT analysis was RR = 0.96 (95% CI 0.6 to 1.54) [28].

Women's Health Study (WHS)

The Women's Health Study (WHS) studied the effects of aspirin vs placebo on major cardiovascular events (Table 1). The main MSM-based result was a RR of 0.93 (95% CI 0.81 to 1.07) [45]. The main ITT effect was a RR of 0.91 (95% CI 0.8 to 1.03) [27].

Ranapurwala

The study by Ranapurwala et al. assessed the effect of a pre-dive checklist on diving mishaps (Table 1). The main MSM-based result was a RR of 0.68 (95% CI 0.5 to 0.93) [40]. The main ITT analysis was a RR of 0.8 (95% CI 0.58 to 1.08) [40].

PREDIMED

The PREDIMED study investigated the effect of Mediterranean diet supplemented with extra-virgin olive oil or nuts versus advice on a low-fat diet on major cardiovascular events (Table 1).MSM-based effects were only reported for the secondary outcome type 2 diabetes mellitus incidence. The MSM-based HR was 0.7 (95% CI 0.52 to 0.94) for extra-virgin olive oil and 0.82 (95% CI 0.62 to 1.09) for nuts versus advice on a low-fat diet [23]. The according main ITT effects were HR 0.69 (95% CI 0.51 to 0.92) and 0.81 (95% CI 0.61 to 1.08), respectively [23].

Women's Health Initiative (WHI)

The Women's Health Initiative (WHI) explored the effect of estrogen-plus-progestin versus placebo in postmenopausal women with intact uterus on coronary heart disease (Table 1). The main MSM-based effect was a HR of 1.69 (95% CI 0.98 to 2.89) [29] and the main ITT effect was a HR of 1.29 (95% CI 1.02 to 1.63) [39].

AIDS Clinical Trial Group 320 study (ACTG 320)

The AIDS Clinical Trial Group 320 study (ACTG 320) compared the effect of zidovudine and lamivudine plus indinavir versus zidovudine and lamivudine alone on AIDS or death in HIV positive, immunosuppressed, ART-experienced patients (Table 1). The main MSM-based result was a HR of 0.45 (95% CI 0.27 to 0.74) [25]. The main ITT result was an HR of 0.5 (95% CI 0.33 to 0.76) [38].

POINTBREAK Study

The POINTBREAK Study assessed the effect of pemetrexed/carboplatin/bevacizumab followed by maintenance pemetrexed/bevacizumab versus maintenance bevacizumab on the overall survival of patients with lung cancer (Table 1). The main MSM-based result was the adaptively truncated MSM (truncating the longitudinal inverse-probability computations) with a HR of 1.12 (95% CI not reported) [46]. The main ITT result was a HR of 1.00 (95% CI 0.86 to 1.16) [43].

Tunis / Faries

The RCT reported by Tunis et al. [44] and Faries et al. [33] studied the effect of olanzapine or risperidone versus a "fail-first" algorithm (conventional antipsychotics then olanzapine if indicated) on change in the brief psychiatric rating scale in patients with schizophrenia or a schizoaffective disorder (Table 1). The main MSM-based estimated treatment difference (olanzapine - conventional) was 1.9 (95% CI 0.5 to 3.3) [33] and the corresponding ITT estimated treatment difference was 0.2 (95% CI -1.8 to 2.1) [44].

Negoro / Yamaguchi

The RCT reported by Negoro et al. [42] and Yamaguchi et al. [32] explored the effect of irinotecan hydrochloride versus cisplatin in patients with lung cancer on overall survival (Table 1). An MSM-based effect was reported for overall survival in the subgroup of patients with stage IIIB lung cancer. The HR was 1.2 (95% CI 0.64 to 2.28) [32]. The main ITT analysis was HR = 1.24 (95% CI 0.81 to 1.91) [42].

CALERIE Study

The CALERIE Study explored the effect of calorie restriction (behavioral approach with dietary modifications) versus no dietary restriction on resting metabolic rate (kcal/d) and core temperature (°C) in healthy young- and middle-aged non-obese men and women (Table 1). We used resting metabolic rate as main comparison. The main MSM-based mean difference was -36 (95% CI not reported) [31] and the corresponding ITT mean difference was -64 (95% CI not reported) [30].

Kisumu RCT

The Kisumu RCT assessed the effect of immediate versus delayed circumcision on HIV incidence (Table 1). The main MSM-based result was a HR of 0.42 (95% CI 0.26 to 0.66) over a 6-year follow-up. There were no ITT-based 6-year follow-up results available (planned duration of the trial was 2 years) [22]. Hence, this trial was not used in the main comparison of MSM vs ITT.

Discussion

Overall findings

In doctoral project I, we found that many observational studies in high impact journals lack satisfactory discussion of confounding bias. When confounding bias was mentioned, authors were often confident that it was rather irrelevant to their findings and they rarely called for cautious interpretation. Studies that discussed possible limitations due to confounding were actually cited more by other researchers than studies that deemed an influence due to confounding unlikely.

In doctoral project II, we found that effects of the tested treatments in non-randomized studies using MSM often pointed in the opposite direction than when RCTs tested the treatments. Overall, MSM-studies tended to show more favorable effects of the experimental treatment; this was more pronounced when the MSM-studies focused on informing health care decision making rather than statistical methodology.

In doctoral project III, we found that MSM was thus far sometimes applied to RCTs, often to adjust for protocol deviations. They were typically not pre-specified, exploratory analyses. Within the main study publication and any corresponding publication reporting MSM-based results, trial authors reported on average 6 analysis results for one clinical question and the spread between the smallest and largest effect estimate can be substantial. The effect estimates of these analyses rarely pointed in different directions. MSM-effects and ITT-effects always pointed in the same direction (i.e. benefit or harm) with MSM-based results being more extreme (i.e. further from the null) in more than half of the cases, and differences in effect sizes were substantial in some cases.

Findings in context

Confounding bias is a pervasive threat to validity of non-randomized studies and deserves utmost attention. The assumption to know all confounding factors, to measure them correctly and to implement them correctly in the statistical models is very unrealistic. In light of the results of this doctoral thesis, major questions can be asked, including “How much impact does confounding have on healthcare decisions?” or “How should healthcare decision makers decide which estimate to trust in?”.

There is no unified answer to the question how much impact confounding has. A simulation study found that the bias in the treatment effect estimate decreases or increases depending on the correlation of the confounders to each other (i.e. whether they consistently bias the effect in the same direction or not), the measurement error, the degree of unmeasured confounding, and the correlation between confounders with the exposure[35]. The impact of unmeasured confounding is greater when the confounders are not correlated as omitting even one confounder from the analysis can lead to substantial bias in the estimated effect[35]. When researchers try to minimize confounding and other biases in observational studies, the size of effect estimates, even the conclusions they draw, may still differ from those they would have obtained in an RCT, even when modern causal models are used (doctoral project II). However, the impact of confounding on the results is rarely discussed in the highest impact literature of general medical, epidemiologic, and specialty journals (doctoral project I)[33]. This is, however, clearly suggested by reporting guidelines, in particular by the Strengthening the Reporting of OBservational studies in Epidemiology (STROBE) guideline. A recent study evaluated the quality of reporting changes after the introduction of STROBE[36]. The authors concluded that despite

some improvements, the reporting quality of confounding remained overall suboptimal[36]. To improve the reporting beyond STROBE, the journal PLOS Medicine requires authors to complete the reporting checklist along with explanations how each item was accomplished[37]. Our findings highlight that such approaches may be urgently needed to improve the reporting of research results.

There are many areas in healthcare where no RCT evidence is available and treatment decisions are made based on observational studies which rely on the assumption of no relevant confounding bias. The current use of observational research is often not clearly focused on such evidence gaps[38]. It is very likely that many health care decisions would be made differently if trials were available. Therefore, confounding bias has probably a huge impact on current health care, in particular in areas where no randomized trial evidence exists.

Even in areas where trials support healthcare decisions, it is not always clear which result of the trial should be used to inform the decision. When multiple analyses (with or without addressing a specific type of confounding) are available for one outcome in an RCT, the spread between the smallest and largest effect estimate can be substantial (doctoral project III). Unclear reporting of multiple effect estimates on one clinical question makes it difficult for clinical decision makers to assess their role for the interpretation of results and to decide which estimate to base their decision on. In addition, due to the complexity of the underlying models, much more details are required to allow for replication of the studies.

After all, depending on the clinical case at hand, even relatively small risk differences may be deciding over life and death. Hence, confounding should always be considered and minimized, if possible on design level[2 39], and reporting needs to be as transparent as possible.

Limitations and future research

Despite applying various rigorous methods, the doctoral projects possess some limitations.

First, in doctoral projects II and III, the analyzed sample was rather small. Our searches indicated that there are no trials for numerous clinical questions explored in observational MSM-studies but not in RCTs, which could be an explanation of the small sample of doctoral project II (as it depended on dyads of observational studies using MSM and RCTs on the same clinical question). The use of MSM in RCTs is relatively new and mostly used for explorative post-hoc analyses, which explains the small sample in doctoral project III. As the merits of this approach for RCTs may become clearer, so may the possible sample increase to allow for future replications or re-analyses.

Second, we conducted the database searches only on PubMed[40]. However, there is no specific recommendation how many and which databases would be necessary to identify a representative sample for meta-epidemiological studies informing on research methodology. We applied a number of methods to improve the validity of our searches: we peer-reviewed the database searches on PubMed, we applied several measures to increase the retrieval rate of relevant studies, and conducted extensive prospective and retrospective citation searches (i.e. studies that have been cited by key literature or are citing key literature)[41].

Third, the research question of doctoral projects II and III was restricted to the application of MSM, and the implications do not necessarily generalize to other models for causal inference. While marginal structural models is one of the most frequently used methods for causal inference and to assess the

impact of time-varying confounding[26], there are other methods to estimate treatment effects in the presence of time-varying confounders including g-computation formula and the g-estimation of structural nested models[26]. Acknowledging a steep increase in the use of these alternative methods, future research could address the agreement of different analysis methods correcting for time-varying confounding and compare it with standard methods.

Fourth, we did not explore mechanistic reasons underlying our findings nor how to tackle some of the discovered issues, i.e. we cannot clearly answer the question why confounding was underestimated in the conclusions of observational research, why the application of MSM to non-randomized data does not always answer research questions as RCTs would, and why many RCTs report several different analysis methods for one and the same result without explanation. These could be addressed in future research, for example by investigating how well authors, editors, and teachers understand the meaning and impact of confounding. An international survey could help to identify the extent of the problem. Second, the results could provide the rationale for a world-wide campaign with the long-term aim of improving research by promoting control for confounding as one of the basic principles of causal inference.

Fifth, two of the three doctoral projects have not been published as journal article yet. However, we plan to (re-) submit in April 2018, also in light of the peer-review comments from the BMJ. However, some of the preliminary results were widely discussed on national and international meetings, including poster presentations and talks on the largest international conference in this area, the Global Evidence Summit of the Cochrane Collaboration, Campbell Collaboration, Guidelines-International Network, International Society for Evidence-based Health Care, and Joanna Briggs Institute in Cape Town, South Africa (September 2017), where I was awarded with the Thomas C. Chalmers award for the work on doctoral project II. In collaboration with my network of researchers from Switzerland, Germany, Austria, Brazil, USA, Canada, and Australia, I plan to further pursue the evaluation of confounding bias in health care and help to better address its implications for healthcare decision-making.

What we can do now

There is plenty of things that can be done right now by users and producers of research evidence. Clinical decision makers should critically appraise the validity of research results and their applicability to the healthcare question at hand. They need to remain cautious when using non-randomized evidence, even when the most modern and complex models are applied. Researchers should acknowledge confounding more carefully in non-randomized and randomized research and when trying to answer questions that go beyond the pragmatic nature of the intention to treat. As we could show, acknowledgement of confounding is not associated with a lower citation impact and researchers do not need to fear to "perish" when publishing their results with adequate calls for caution. Before launching a trial, researchers should think critically about what they want to measure (including confounding variables), how they want to measure it and to ensure high quality data collection, and how bias, non-adherence, and loss-to-follow-up can be avoided. Adherence to a reporting guideline helps avoiding basic errors of reporting. Useful tools already exist, e.g. over 200 guidelines listed on the EQUATOR Network website (www.equator-network.org). With regard to current results, researchers should clearly state which analyses were pre-specified and which were explorative. Journals should make their policies clear and adhere to them.

Closing Remarks

"Knowledge is merely brilliance in organization of ideas and not wisdom. The truly wise person goes beyond knowledge." – Confucius[42]

One issue with knowledge is the things we do not know. Semmelweis was laughed at, ignored at best, for his idea that organic particles from the autopsy room could cause the death of subsequently treated patients, especially women giving birth[15]. Without knowing the causal pathways, his hygienic measures had such a strong effect that eventually he convinced his peers of properly disinfecting their hands with chlorine solution before attending a delivery – with great life-saving effects. Yet, his explanation seemed too fantastic for the medical community and he received strong opposition[15]. Today, we know that these "particles" were bacteria and we found the means to see and measure them. Yet, the history of confounding is quickly forgotten when researchers succumb to the pressure of the current publication-based reward system rather than trying to evoke an overdue paradigm shift. This dissertation is a reminder to look beyond, to be realistic about the things we do not know, and to be open for the possibility that we got it all wrong: Do not make science an alternative fact.

References

1. Sackett DL, Rosenberg WM, Gray JA, Haynes RB, Richardson WS. Evidence based medicine: what it is and what it isn't. 1996. Clinical orthopaedics and related research 2007;**455**:3-5.
2. Hannan EL. Randomized clinical trials and observational studies: guidelines for assessing respective strengths and limitations. JACC Cardiovasc Interv 2008;**1**(3):211-7.
3. Higgins JPT, Green S. Cochrane Handbook for Systematic Reviews of Interventions Version 5.1.0 [updated March 2011]. The Cochrane Collaboration 2011;Available from www.cochrane-handbook.org.
4. Greenland S. Multiple-bias modelling for analysis of observational data. Journal of the Royal Statistical Society: Series A (Statistics in Society) 2005;**168**(2):267-306.
5. Chavalarias D, Ioannidis JP. Science mapping analysis characterizes 235 biases in biomedical research. Journal of clinical epidemiology 2010;**63**(11):1205-15.
6. Mansournia MA, Higgins JP, Sterne JA, Hernan MA. Biases in Randomized Trials: A Conversation Between Trialists and Epidemiologists. Epidemiology (Cambridge, Mass) 2017;**28**(1):54-59.
7. Grimes DA, Schulz KF. Bias and causal associations in observational research. Lancet 2002;**359**(9302):248-52.
8. Hernan MA, Hernandez-Diaz S, Robins JM. A structural approach to selection bias. Epidemiology (Cambridge, Mass) 2004;**15**(5):615-25.
9. Kahan BC, Rehal S, Cro S. Risk of selection bias in randomised trials. Trials 2015;**16**:405.
10. Vandenbroucke JP. The history of confounding. Sozial- und Präventivmedizin 2002;**47**(4):216-24.
11. Kyriacou DN, Lewis RJ. Confounding by Indication in Clinical Research. Jama 2016;**316**(17):1818-19.
12. Braga LHP, Farrokhyar F, Bhandari M. Practical Tips for Surgical Research: Confounding: What is it and how do we deal with it? Canadian Journal of Surgery 2012;**55**(2):132-38.
13. Christenfeld NJ, Sloan RP, Carroll D, Greenland S. Risk factors, confounding, and the illusion of statistical control. Psychosom Med 2004;**66**(6):868-75.
14. Cornfield J, Haenszel W, Hammond EC, Lilienfeld AM, Shimkin MB, Wynder EL. Smoking and lung cancer: recent evidence and a discussion of some questions. International Journal of Epidemiology 2009;**38**(5):1175-91.
15. Nuland SB. *The Doctors' Plague Germs, Childbed Fever, and the Strange Story of Ignac Semmelweis*: Paw Prints, 2008.
16. Armstrong K. Methods in comparative effectiveness research. Journal of clinical oncology : official journal of the American Society of Clinical Oncology 2012;**30**(34):4208-14.
17. Mansournia MA, Higgins JPT, Sterne JAC, Hernán MA. Biases in randomized trials: a conversation between trialists and epidemiologists. Epidemiology (Cambridge, Mass) 2017;**28**(1):54-59.
18. Karanicolas PJ, Montori VM, Devereaux PJ, Schunemann H, Guyatt GH. A new 'mechanistic-practical" framework for designing and interpreting randomized trials. Journal of clinical epidemiology 2009;**62**(5):479-84.
19. Hernan MA, Robins JM. Per-Protocol Analyses of Pragmatic Trials. The New England journal of medicine 2017;**377**(14):1391-98.
20. Toh S, Hernan MA. Causal inference from longitudinal studies with baseline randomization. Int J Biostat 2008;**4**(1):Article 22.
21. Altman DG, Bland JM. Statistics notes: Absence of evidence is not evidence of absence. BMJ (Clinical research ed) 1995;**311**(7003):485.
22. Young JG, Hernan MA, Picciotto S, Robins JM. Relation between three classes of structural models for the effect of a time-varying exposure on survival. Lifetime Data Anal 2010;**16**(1):71-84.
23. Suarez D, Borras R, Basagana X. Differences between marginal structural models and conventional models in their exposure effect estimates: a systematic review. Epidemiology (Cambridge, Mass) 2011;**22**(4):586-8.

24. Kurth T, Walker AM, Glynn RJ, Chan KA, Gaziano JM, Berger K, Robins JM. Results of multivariable logistic regression, propensity matching, propensity adjustment, and propensity-based weighting under conditions of nonuniform effect. American journal of epidemiology 2006;**163**(3):262-70.
25. Robins JM, Hernan MA, Brumback B. Marginal structural models and causal inference in epidemiology. Epidemiology (Cambridge, Mass) 2000;**11**(5):550-60.
26. Williamson T, Ravani P. Marginal structural models in clinical research: when and how to use them? Nephrology, dialysis, transplantation : official publication of the European Dialysis and Transplant Association - European Renal Association 2017;**32**(suppl_2):ii84-ii90.
27. Hernán MA, Brumback B, Robins JM. Marginal Structural Models to Estimate the Joint Causal Effect of Nonrandomized Treatments. Journal of the American Statistical Association 2001;**96**(454):440-48.
28. Rehkopf DH, Glymour MM, Osypuk TL. The Consistency Assumption for Causal Inference in Social Epidemiology: When a Rose is Not a Rose. Current epidemiology reports 2016;**3**(1):63-71.
29. Lawlor DA, Davey Smith G, Kundu D, Bruckdorfer KR, Ebrahim S. Those confounded vitamins: what can we learn from the differences between observational versus randomised trial evidence? Lancet 2004;**363**(9422):1724-7.
30. Grodstein F, Stampfer MJ, Manson JE, Colditz GA, Willett WC, Rosner B, Speizer FE, Hennekens CH. Postmenopausal estrogen and progestin use and the risk of cardiovascular disease. The New England journal of medicine 1996;**335**(7):453-61.
31. Zhang Z. Meta-epidemiological study: a step by step approach by using R. J Evid Based Med 2016.
32. Ioannidis JPA, Patsopoulos NA, Rothstein HR. Reasons or excuses for avoiding meta-analysis in forest plots. BMJ : British Medical Journal 2008;**336**(7658):1413-15.
33. Hemkens LG, Ewald H, Naudet F, Ladanie A, Shaw JG, Sajeev G, Ioannidis JPA. Interpretation of epidemiologic studies very often lacked adequate consideration of confounding. Journal of clinical epidemiology 2017.
34. Hemkens LG, Ewald H, Naudet F, Ladanie A, Shaw JG, Sajeev G, Ioannidis JPA. Interpretation of epidemiologic studies very often lacked adequate consideration of confounding. Journal of Clinical Epidemiology 2018;**93**:94-102.
35. Fewell Z, Davey Smith G, Sterne JA. The impact of residual and unmeasured confounding in epidemiologic studies: a simulation study. American journal of epidemiology 2007;**166**(6):646-55.
36. Pouwels KB, Widyakusuma NN, Groenwold RH, Hak E. Quality of reporting of confounding remained suboptimal after the STROBE guideline. Journal of clinical epidemiology 2016;**69**:217-24.
37. The PME. Observational Studies: Getting Clear about Transparency. PLoS medicine 2014;**11**(8):e1001711.
38. Hemkens LG, Contopoulos-Ioannidis DG, Ioannidis JP. Routinely collected data and comparative effectiveness evidence: promises and limitations. CMAJ : Canadian Medical Association journal = journal de l'Association medicale canadienne 2016;**188**(8):E158-64.
39. Tunis SR, Stryer DB, Clancy CM. Practical clinical trials: increasing the value of clinical research for decision making in clinical and health policy. Jama 2003;**290**(12):1624-32.
40. McDonagh M, Peterson K, Raina P, Chang S, Shekelle P. Methods Guide for Effectiveness and Comparative Effectiveness Reviews. Rockville (MD): Agency for Healthcare Research and Quality (US); 2008-. Available from: https://www.ncbi.nlm.nih.gov/books/NBK47095/. Secondary Methods Guide for Effectiveness and Comparative Effectiveness Reviews. Rockville (MD): Agency for Healthcare Research and Quality (US); 2008-. Available from: https://www.ncbi.nlm.nih.gov/books/NBK47095/ 2013.
41. Cooper C, Booth A, Britten N, Garside R. A comparison of results of empirical studies of supplementary search techniques and recommendations in review methodology handbooks: a methodological review. Syst Rev 2017;**6**(1):234.
42. Students' Academy. *Words of Wisdom: Confucius*: Lulu.com, 2014.

Appendix I – Further Manuscripts published during doctoral studies

Systematic review and simulation study of ignoring clustered data in surgical trials

Br J Surg. 2018;105:182-91.

Dell-Kuster S, Droeser RA, Schafer J, Gloy V, Ewald H, Schandelmaier S, Hemkens LG, Bucher HC, Young J, Rosenthal R

Background: Multiple surgical procedures in a single patient are relatively common and lead to dependent (clustered) data. This dependency needs to be accounted for in study design and data analysis. A systematic review was performed to assess how clustered data were handled in inguinal hernia trials. The impact of ignoring clustered data was estimated using simulations.

Methods: PubMed, Embase and the Cochrane Library were reviewed systematically for RCTs published between 2004 and 2013, including patients undergoing unilateral or bilateral inguinal hernia repair. Study characteristics determining the appropriateness of handling clustered data were extracted. Using simulations, various statistical methods accounting for clustered data were compared with an analysis ignoring clustering by assuming 100 hernias, with a varying percentage of patients having bilateral hernias.

Results: Of the 50 eligible trials including patients with bilateral hernias, 20 (40 per cent) did not provide information on how they dealt with clustered data and 18 (36 per cent) avoided clustering by assessing the outcome by patient and not by hernia. None of the remaining 12 trials (24 per cent) considered clustering in the design or analysis. In the simulations, ignoring clustering led to an increased type I error rate of up to 12 per cent and to a loss in power of up to 15 per cent, depending on whether the patient or the hernia was the randomization unit.

Conclusion: Clustering was rarely considered in inguinal hernia trials. The simulations underline the importance of considering clustering as part of the statistical analysis to avoid false-positive and false-negative results, and hence inappropriate study conclusions.

Off-label treatments were not consistently better or worse than approved drug treatments in randomized trials

J Clin Epidemiol. 2017.

Ladanie A, Ioannidis JPA, Stafford RS, Ewald H, Bucher HC, Hemkens LG.

Objectives: Off-label drug use is highly prevalent but controversial and often discouraged assuming generally inferior medical effects associated with off-label use.

Study Design and Setting: We searched PubMed, MEDLINE, PubMed Health, and the Cochrane Library up to May 2015 for systematic reviews including meta-analyses of randomized clinical trials (RCTs) comparing off-label and approved drugs head-to-head in any population and on any medical outcome. We combined the comparative effects in meta-analyses providing summary odds ratios (sOR) for each treatment comparison and outcome, and then calculated an overall summary of the sOR across all comparisons (ssOR).

Results: We included 25 treatment comparisons with 153 RCTs and 24,592 patients. In six of 25 comparisons (24%), off-label drugs were significantly superior (five of 25) or inferior (one of 25) to approved treatments. There was substantial statistical heterogeneity across comparisons (I2 = 43%). Overall, off-label drugs were more favorable than approved treatments (ssOR 0.72; 95% CI = 0.54–0.95). Analyses of patient-relevant outcomes were similar (statistical significant differences in 24% (six of 25); ssOR 0.74; 95% CI = 0.56–0.98; I2 = 60%). Analyses of primary outcomes of the systematic reviews (n = 22 comparisons) indicated less heterogeneity and no statistically significant difference overall (ssOR 0.85; 95% CI = 0.67–1.06; I2 = 0%).

Conclusion: Approval status does not reliably indicate which drugs are more favorable in situations with clinical trial evidence comparing off-label with approved use. Drug effectiveness assessments without considering off-label use may provide incomplete information. To ensure that patients receive the best available care, funding, policy, reimbursement, and treatment decisions should be evidence based considering the entire spectrum of available therapeutic choices.

Comparative effectiveness of tenofovir in HIV-infected treatment-experienced patients: systematic review and meta-analysis

HIV Clin Trials. 2016:1-11.

Ewald H, Santini-Oliveira M, Buhler JE, Vuichard D, Schandelmaier S, Stöckle M, Briel M, Bucher HC, Hemkens LG

Background: Antiretroviral therapy (ART) regimens for HIV infection are frequently changed. We conducted a systematic review of randomized trials (RCTs) on the benefits and harms of switching to tenofovir disoproxil fumarate (TDF)-based regimens in ART-experienced patients.

Methods: We included RCTs in HIV-infected adults comparing switching to a TDF-containing regimen with maintaining or switching to another regimen. We searched MEDLINE, EMBASE, CENTRAL, LILACS, SCI, and the WHO Global Health Library. We assessed bias with the Cochrane tool and synthesized data using random-effects meta-analyses and Peto's approach. For further analyses, we added data from a previous systematic review in treatment-naïve patients.

Results: 17 RCTs with 2210 patients were included. All but one study had a high risk of bias. There was no significant association of switching to TDF-based regimens with mortality, fractures, CD4-cell count, body fat, virological failure, LDL-, and HDL-cholesterol. TDF-based regimens decreased total cholesterol (mean difference −12.05 mg/dL; 95% CI −20.76 to −3.34), trigylcerides (−14.33 mg/dL; −23.73 to −4.93), and bone mineral density (BMD; hip: −2.46%; −3.9 to −1.03; lumbar spine −1.52%; −2.69 to −0.34). Effects on estimated glomerular filtration (eGFR) were inconsistent and depended on the measurement. Adding 22 RCTs from 8297 treatment-naïve patients gave consistent results with then significant reductions of LDL (−7.57 mg/dL; −10.37 to −4.78), HDL (−2.38 mg/dL; −3.83 to −0.93), and eGFR (−3.49 ml/min; −5.56 to −1.43).

Conclusions: Switching to TDF-based regimens is associated with reductions of BMD and lipid levels and possibly lowered kidney function. The evidence is limited by the high risk of bias.

Colchicine and prevention of cardiovascular events

JAMA. 2016;316(10):1106-1107.

Hemkens LG, Ewald H, Briel M.

JAMA Clinical Evidence Synopsis, no abstract available.

The clinical effectiveness of pneumococcal conjugate vaccines – a systematic review and meta-analysis of randomized controlled trials

Dtsch Arztebl Int 2016; 113(9): 139-46; DOI: 10.3238/arztebl.2016.0139.

Ewald H, Briel M, Vuichard D, Kreutle V, Zhydkov A, Gloy V.

(English)

Background: Streptococcus pneumoniae is responsible for approximately 1.6 million yearly deaths worldwide. An up-to-date evidence base on the effects of pneumococcal conjugate vaccines (PCVs) on infectious diseases and mortality in any population or setting regardless of age or health status is currently lacking.

Methods: We systematically searched MEDLINE and EMBASE for pertinent randomized controlled trials (RCTs). Two reviewers independently screened 9498 titles/abstracts and 430 full-text papers for eligible trials. The outcomes of our meta-analysis were pooled using relative risks (RRs) with a random effects model or Peto's odds ratios (ORs) if event rates were <1%.

Results: 21 RCTs comprising 361 612 individuals were included. PCVs reduced the risk for invasive pneumococcal disease (odds ratio [OR]: 0.43, 95% confidence interval [CI]: [0.36; 0.51]), all-cause acute otitis media (AOM) (RR: 0.93, 95% CI: [0.86; 1.00]), pneumococcal AOM (RR: 0.57, 95% CI: [0.39; 0.83]), all-cause pneumonia (RR: 0.93, 95% CI: [0.89; 0.97]), and pneumococcal pneumonia (RR: 0.78, 95% CI: [0.62; 0.97]). We found no significant effect of PCVs on all-cause mortality (RR: 0.95, 95% CI: [0.88; 1.03]) or recurrent AOM (RR: 0.87, 95% CI: [0.72; 1.05]).

Conclusion: PCVs are associated with large risk reductions for pneumococcal infectious diseases, smaller risk reductions for infectious diseases from any cause, and no significant effect on all-cause mortality.

Klinische Wirksamkeit von Pneumokokken-Konjugatimpfstoffen – Systematische Übersichtsarbeit und Metaanalyse randomisierter kontrollierter Studien

(German adaptation)

Hintergrund: Streptococcus pneumoniae ist jährlich für rund 1,6 Millionen Todesfälle weltweit verantwortlich. Derzeit gibt es keine systematische Übersichtsarbeit zu Pneumokokken-Konjugatimpfstoffen (PCV), in der die Wirksamkeit in Bezug auf die Reduktion von Infektionskrankheiten und der Gesamtsterblichkeit in verschiedenen Populationen und Settings beurteilt wird.

Methoden: Im Rahmen einer systematischen Literaturrecherche wurden MEDLINE und Embase nach geeigneten randomisierten kontrollierten Studien (RCT) durchsucht. Zwei Autoren überprüften unabhängig voneinander 9 498 Titel/Abstracts und 430 Volltexte auf relevante Studien. Die Ergebnisse der Metaanalyse wurden als relative Risiken (RR) mit Random-Effects-Modellen, und bei Ereignisraten unter 1 % als Peto's Odds Ratios (OR) dargestellt.

Ergebnisse: Es konnten 21 RCT mit 361 612 Personen eingeschlossen werden. PCV reduzierten das Risiko für invasive Pneumokokken-Erkrankungen (OR: 0,43; 95-%-Konfidenzintervall [0,36; 0,51]), akute Otitis media (RR: 0,93 [0,86; 1,00]), Pneumokokken-spezifische akute Otitis media (RR: 0,57 [0,39; 0,83]), Pneumonie (RR: 0,93 [0,89; 0,97]) und Pneumokokken-spezifische Pneumonie (RR: 0,78 [0,62; 0,97]). Es zeigte sich kein signifikanter Effekt von PCV auf die Gesamtsterblichkeit (RR: 0,95 [0,88; 1,03]) oder auf rezidivierende akute Otitis media (RR: 0,87 [0,72; 1,05]).

Schlussfolgerung: Pneumokokken-Konjugatimpfstoffe sind mit großen Risikoreduktionen für Pneumokokken-spezifische Infektionskrankheiten, kleineren Risikoreduktionen für Infektionskrankheiten jedweder Ursache und keinem signifikanten Effekt auf die Gesamtsterblichkeit assoziiert.

Cardiovascular effects and safety of long-term colchicine treatment: Cochrane review and meta-analysis

Heart 2016;0:1–7.

Hemkens LG, Ewald H, Gloy VL, Arpagus A, Olu KK, Nidorf M, Glinz D, Nordmann AJ, Briel M.

Colchicine is an old anti-inflammatory drug that has shown substantial cardiovascular benefits in recent trials. We systematically reviewed cardiovascular benefits and harms of colchicine in any population and specifically in patients with high cardiovascular risk. We evaluated randomised controlled trials comparing colchicine over at least 6 months versus any control in any adult population. Primary outcomes were all-cause mortality, myocardial infarction and adverse events. Cardiovascular mortality was a secondary outcome. We included 39 trials with 4992 patients. The quality of evidence for mortality outcomes and myocardial infarction was moderate but lower for adverse events. Colchicine had no effect on all-cause mortality (RR 0.94, 95% CI 0.82 to 1.09; I^2=27%; 0 trials). Cardiovascular mortality was reduced in some but not all meta-analytical models (random-effects RR 0.34, 0.09 to 1.21, I^2=9%; Peto's OR 0.24, 0.09 to 0.64, I^2=15%; Mantel-Haenszel fixed-effect RR 0.20, 0.06 to 0.68, I^2=0%; 7 trials). The risk for myocardial infarction was reduced (RR 0.20, 0.07 to 0.57; 2 trials). There was no effect on total adverse events (RR 1.52, 0.93 to 2.46, I^2=45%; 11 trials) but gastrointestinal intolerance was increased (RR 1.83, 1.03 to 3.26, I^2=74%; 11 trials). Reporting of serious adverse events was inconsistent; no event occurred over 824 patient-years (4 trials). Effects in high cardiovascular risk populations were similar (4 trials; 1230 patients). We found no evidence supporting colchicine doses above 1 mg/day. Colchicine may have substantial cardiovascular benefits; however, there is sufficient uncertainty about its benefit and harm to indicate the need for large-scale trials to further evaluate this inexpensive, promising treatment in cardiovascular disease.

Colchicine for prevention of cardiovascular events

Cochrane Database Syst Rev. 2016;1:CD011047.

Hemkens LG, Ewald H, Gloy VL, Arpagus A, Olu KK, Nidorf M, Glinz D, Nordmann AJ, Briel M.

Background: Colchicine is an anti-inflammatory drug that is used for a wide range of inflammatory diseases. Cardiovascular disease also has an inflammatory component but the effects of colchicine on cardiovascular outcomes remain unclear. Previous safety analyses were restricted to specific patient populations.

Objectives: To evaluate potential cardiovascular benefits and harms of a continuous long-term treatment with colchicine in any population, and specifically in people with high cardiovascular risk.

Search methods: We searched the Cochrane Central Register of Controlled Trials (CENTRAL), MEDLINE, EMBASE, ClinicalTrials.gov, WHO International Clinical Trials Registry, citations of key papers, and study references in January 2015. We also contacted investigators to gain unpublished data.

Selection criteria: Randomised controlled trials (parallel-group or cluster design or first phases of cross-over studies) comparing colchicine over at least six months versus any control in any adult population.

Data collection and analysis: Primary outcomes were all-cause mortality, myocardial infarction, and adverse events. Secondary outcomes were cardiovascular mortality, stroke, heart failure, non-scheduled hospitalisations, and non-scheduled cardiovascular interventions. We conducted predefined subgroup analyses, in particular for participants with high cardiovascular risk.

Main results: We included 39 randomised parallel-group trials with 4992 participants. Colchicine had no effect on all-cause mortality (RR 0.94, 95% CI 0.82 to 1.09; participants = 4174; studies = 30; I^2 = 27%; moderate quality of evidence). There is uncertainty surrounding the effect of colchicine in reducing cardiovascular mortality (RR 0.34, 95% CI 0.09 to 1.21, I^2 = 9%; participants = 1132; studies = 7; moderate quality of evidence). Colchicine reduced the risk for total myocardial infarction (RR 0.20, 95% CI 0.07 to 0.57; participants = 652; studies = 2; moderate quality of evidence). There was no effect on total adverse events (RR 1.52, 95% CI 0.93 to 2.46; participants = 1313; studies = 11; I^2 = 45%; very low quality of evidence) but gastrointestinal intolerance was increased (RR 1.83, 95% CI 1.03 to 3.26; participants = 1258; studies = 11; I^2 = 74%; low quality of evidence). Colchicine showed no effect on heart failure (RR 0.62, 95% CI 0.10 to 3.88; participants = 462; studies = 3; I^2 = 45%; low quality of evidence) and no effect on stroke (RR 0.38, 95% CI 0.09 to 1.70; participants = 874; studies = 3; I^2 = 45%; low quality of evidence). Reporting of serious adverse events was inconsistent; no event occurred over 824 patient-years (4 trials). Effects on other outcomes were very uncertain. Summary effects of RCTs specifically focusing on participants with high cardiovascular risk were similar (4 trials; 1230 participants).

Authors' conclusions: There is much uncertainty surrounding the benefits and harms of colchicine treatment. Colchicine may have substantial benefits in reducing myocardial infarction in selected high-risk populations but uncertainty about the size of the effect on survival and other cardiovascular outcomes is high, especially in the general population from which most of the studies in our review were drawn. Colchicine is associated with gastrointestinal side effects based on low-quality evidence. More evidence from large-scale randomized trials is needed.

Comparative effectiveness of Tenofovir in treatment-naïve HIV-infected patients: systematic review and meta-analysis

HIV Clin Trials. 2015 Oct;16(5):178-89.

Hemkens LG, Ewald H, Santini-Oliveira M, Bühler J-E, Vuichard D, Schandelmaier S, Stöckle M, Briel M, Bucher HC.

Introduction: Benefits and harms of tenofovir disoproxil fumarate (TDF) in HIV-infected, antiretroviral treatment (ART)-naïve patients of any age have not been systematically reviewed since recent milestone trials were published.

Methods: We searched MEDLINE, EMBASE, CENTRAL, SCI, LILACS, WHO GHL, and ClinicalTrials.gov for randomized controlled trials (RCTs) comparing TDF-based treatments with any other ART-regimen (last search 01/2015). Trial characteristics and results were extracted, risks of bias systematically assessed, and treatment effects synthesized in meta-analyses using random-effects models.

Results: We included 22 RCTs (8297 patients). We found no differences between groups for mortality, AIDS, fractures, CD4 cell count, and virological failure; and inconclusive information due to inadequate reporting for cardiovascular events, renal failure, proteinuria, rash, and quality of life. Tenofovir disoproxil fumarate-based regimens significantly reduced total cholesterol (mean difference { 18.42 mg/dl; 95% confidence interval [CI] { 22.80 to { 14.0), LDL-cholesterol ({ 9.53 mg/dl; { 12.16 to { 6.89), HDL-cholesterol ({ 2.97 mg/dl; { 4.41 to { 1.53), and triglycerides ({ 29.77 mg/dl; { 38.61 to { 20.92), bone mineral density (BMD) (hip: { 1.41%; { 1.87 to { 0.94), and glomerular filtration rate (eGFR) ({ 3.47 ml/minute; { 5.89 to { 1.06) over 48 weeks of follow-up. Effects were similar in trials comparing fixed-dose TDF/FTC-based regimens with ABC/3TC-based regimens. We found no influence of baseline viral load on virological failure.

Discussion: Moderate-quality evidence suggests similar effects of TDF-based treatment regimens and other ART on virological failure. Tenofovir disoproxil fumarate-based regimens are associated with a more favorable lipid profile, but with increased risk of reduced BMD and eGFR. Improved reporting quality is vital to allow assessment of clinical outcomes in future trials.

Adjunctive corticosteroids for Pneumocystis jiroveci pneumonia in patients with HIV infection

Cochrane Database of Systematic Reviews 2015, Issue 4. Art. No.: CD006150. DOI: 10.1002/14651858.CD006150.pub2.

Ewald H, Raatz H, Boscacci R, Furrer H, Bucher HC, Briel M.

Background: Pneumocystis jiroveci pneumonia (PCP) remains the most common opportunistic infection in patients infected with the human immunodeficiency virus (HIV). Among patients with HIV infection and PCP the mortality rate is 10% to 20% during the initial infection and this increases substantially with the need for mechanical ventilation. It has been suggested that corticosteroids adjunctive to standard treatment for PCP could prevent the need for mechanical ventilation and decrease mortality in these patients.

Objectives: To assess the effects of adjunctive corticosteroids on overall mortality and the need for mechanical ventilation in HIV-infected patients with PCP and substantial hypoxaemia (arterial oxygen partial pressure < 70 mmHg or alveolar-arterial gradient > 35 mmHg on room air).

Search methods: For the original review we searched The Cochrane Library (2004, Issue 4), MEDLINE (January 1980 to December 2004) and EMBASE (January 1985 to December 2004) without language restrictions. We further reviewed the reference lists from previously published overviews, searched UptoDate version 2005 and Clinical Evidence Concise (Issue 12, 2004), contacted experts in the field and searched the reference lists of identified publications for citations of additional relevant articles. In this update of our review, we searched the above-mentioned databases in September 2010 and April 2014 for trials published since our original review. We also searched for ongoing trials in ClinicalTrials.gov and the World Health Organization International Clinical Trial Registry Platform (ICTRP). We searched for conference abstracts via AEGIS.

Selection criteria: Randomised controlled trials that compared corticosteroids to placebo or usual care in HIV-infected patients with PCP in addition to baseline treatment with trimethoprim-sulfamethoxazole, pentamidine or dapsone-trimethoprim, and reported mortality data. We excluded trials in patients with no or mild hypoxaemia (arterial oxygen partial pressure > 70 mmHg or an alveolar-arterial gradient <35 mmHg on room air) and trials with a follow-up of less than 30 days.

Data collection and analysis: Two teams of review authors independently evaluated the methodology and extracted data from each primary study. We pooled treatment effects across studies and calculated a weighted average risk ratio of overall mortality in the treatment and control groups using a random-effects model. In this update of our review, we used the GRADE methodology to assess evidence quality.

Main results: Of 2029 screened records, we included seven studies in the review and six in the meta-analysis. Risk of bias varied: the randomization and allocation process was often not clearly described, five of seven studies were double-blind and there was almost no missing data. The quality of the evidence for mortality was high. Risk ratios for overall mortality for adjunctive corticosteroids were 0.56 (95% confidence interval (CI) 0.32 to 0.98) at one month and 0.59 (95% CI 0.41 to 0.85) at three to four months of follow-up. In adults, to prevent one death, numbers needed to treat are nine patients in a setting without highly active antiretroviral therapy (HAART) available, and 23 patients with HAART

available. The three largest trials provided moderate quality data on the need for mechanical ventilation, with a risk ratio of 0.38 (95% CI 0.20 to 0.73) in favour of adjunctive corticosteroids. One study was conducted in infants, suggesting a risk ratio for death in hospital of 0.81 (95% CI 0.51 to 1.29; moderate quality evidence).

Authors' conclusions: The number and size of trials investigating adjunctive corticosteroids for HIV-infected patients with PCP is small, but the evidence from this review suggests a beneficial effect for adult patients with substantial hypoxaemia. There is insufficient evidence on the effect of adjunctive corticosteroids on survival in infants.

Appendix II – Short curriculum vitae: Hannah Ewald

Education

2015 – present	**PhD in Epidemiology** at University of Basel and Swiss Tropical and Public Health Institute
2011 – 2012	**Master in Public Health** at University of Warwick (Department of Medicine), UK
2008 – 2009	**Bachelor of Health in Physiotherapy** at Hogeschool van Arnhem en Nijmegen, Netherlands
2005 – 2008	**Qualification as state-approved physiotherapist** at Helmut Rödler Schule für Physiotherapie, Chemnitz, Germany
1991 – 2004	**Abitur** at Gregor-Mendel-Gymnasium, Amberg, Germany

Professional Experience

2017 – present	**Information Specialist** at University Library Basel, Switzerland
2013 – present	**Researcher** at Basel Institute for Clinical Epidemiology & Biostatistics, University Hospital Basel, Switzerland
2008 – 2011	**Physio- and Sports-therapist** in Amberg and Erlangen, Germany

Scientific Awards

2017	**Thomas C. Chalmers Award** for best poster presentation on the Global Evidence Summit in Cape Town, South Africa

List of conferences with presentations

2018	**19th Annual Meeting of the German Network for Evidence-based Medicine (DNEbM):** „Wie verändern sich die Hauptergebnisse von Systematic Reviews durch weniger aufwendige Literatursuchen?“ Eine meta-epidemiologische Analyse“ (poster presentation), Graz, Austria
2017	**18th Annual Meeting of the German Network for Evidence-based Medicine (DNEbM):** “Cardiovascular benefits and potential harm of colchicine – a Cochrane review on a new indication of an old drug” (oral presentation), Hamburg, Germany
2017	**Global Evidence Summit 2017:** “Using Evidence. Improving lives” (poster presentation), Cape Town, South Africa
2015	**83. SGIM annual meeting 2015:** “Colchicine for prevention of cardiovascular events: A systematic review and meta-analysis” (oral presentation), Basel, Switzerland

List of teaching activities

2015 – present	**Teaching medical students** in literature search methods and critical appraisal during POEM (Patient-oriented and evidence-based medicine) at University of Basel, Switzerland
2017 – 2018	**Assistance and support with systematic review methodology** to new PhD student at Basel Institute for Clinical Epidemiology & Biostatistics, University Hospital Basel, Switzerland
2015 – 2017	**Assistance and support with systematic review methodology** to Master students at Basel Institute for Clinical Epidemiology & Biostatistics, University Hospital Basel, Switzerland