Journal of Clinical Epidemiology 107 (2019) 12−26

ORIGINAL ARTICLE

Marginal structural models and other analyses allow multiple estimates of treatment effects in randomized clinical trials: Meta-epidemiological analysis

Hannah Ewald[a,b,c], Benjamin Speich[a], Aviv Ladanie[a,b], Heiner C. Bucher[a], John P.A. Ioannidis[d,e,f,g,h], Lars G. Hemkens[a,*]

[a] *Basel Institute for Clinical Epidemiology and Biostatistics, Department of Clinical Research, University Hospital Basel, University of Basel, Basel 4031, Switzerland*

[b] *Swiss Tropical and Public Health Institute, Basel 4051, Switzerland*

[c] *University Medical Library, University of Basel, Basel 4051, Switzerland*

[d] *Stanford Prevention Research Center, Department of Medicine, Stanford University School of Medicine, Stanford, CA 94305, USA*

[e] *Meta-Research Innovation Center at Stanford (METRICS), Stanford University, Palo Alto, CA 94305, USA*

[f] *Department of Health Research and Policy, Stanford University School of Medicine, Stanford, CA 94305, USA*

[g] *Department of Biomedical Data Science, Stanford University School of Medicine, Stanford, CA 94305, USA*

[h] *Department of Statistics, Stanford University School of Humanities and Sciences, Stanford, CA 94305, USA*

Accepted 5 November 2018; Published online 10 November 2018

Abstract

Objectives: To determine how marginal structural models (MSMs), which are increasingly used to estimate causal effects, are used in randomized clinical trials (RCTs) and compare their results with those from intention-to-treat (ITT) or other analyses.

Study Design and Setting: We searched PubMed, Scopus, citations of key references, and Clinicaltrials.gov. Eligible RCTs reported clinical effects based on MSMs and at least one other analysis.

Results: We included 12 RCTs reporting 138 analyses for 24 clinical questions. In 19/24 (79%), MSM-based and other effect estimates were all in the same direction, 22/22 had overlapping 95% confidence intervals (CIs), and in 19/22 (86%), the MSM effect estimate lay within all 95% CIs of all other effects (in two cases no CIs were reported). For the same clinical question, the largest effect estimate from any analysis was 1.19-fold (median; interquartile range 1.13-1.34) larger than the smallest. All MSM and ITT effect estimates were in the same direction and had overlapping 95% CIs. In 71% (12/17), they also agreed on the presence of statistical significance. MSM-based effect estimates deviated more from the null than those based on ITT ($P = 0.18$). The effect estimates of both approaches differed 1.12-fold (median; interquartile range 1.02-1.22).

Data sharing: No additional data available.

Conflict of interest: All authors have completed the Unified Competing Interest form at www.icmje.org/coi_disclosure.pdf. All authors declare no financial relationships with any organizations that might have an interest in the submitted work in the previous 3 years and no other relationships or activities that could appear to have influenced the submitted work.

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. Benjamin Speich is supported by the Research Foundation of the University of Basel. 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.

Clinical trial registration: Not required.

* Corresponding author. Basel Institute for Clinical Epidemiology & Biostatistics, University Hospital Basel, Spitalstrasse 12, Basel CH-4031, Tel.: +41 61 32 85404; fax +41 61 26 53109.

E-mail address: Lars.Hemkens@usb.ch (L.G. Hemkens).

https://doi.org/10.1016/j.jclinepi.2018.11.001

Conclusions: MSMs provided largely similar effect estimates as other available analyses. Nevertheless, some of the differences in effect estimates or statistical significance may become important in clinical decision-making, and the multiple estimates require utmost attention of possible selective reporting bias.

Keywords: Marginal structural models; Intention-to-treat; Randomized controlled trial; Reporting; Vibration of effects; Selective reporting; Prespecification

1. Introduction

Randomized clinical trials (RCTs) are usually the best way to estimate 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 lifestyle intervention, or a treatment that 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 trial protocol. Randomization makes confounding random, and this hopefully improves the chances of getting valid estimates of effects of such decisions [3−5]. Such trials focusing on health care decision-making are often pragmatic or practical [4]. Explanatory or mechanistic trials aim to better understand the underlying causal pathways of the decisions, such as biological mechanisms of treatments. For such research questions, PP analyses (evaluating only patients who adhere to their assigned treatment and the clinical trial instructions as defined in the study protocol [2]) or AT analyses (evaluating patients according to the treatment they received, not the treatment they were assigned to [2]) may be of specific interest. They are often used to estimate the PP effect, i.e., "the causal effect of treatment that would have been observed if all individuals had adhered to their assigned treatment as specified in the protocol of the experiment" [6].

The conceptual difference of ITT and PP effects (or estimands) has gained more attention for clinical trial design recently, e.g., through the ICH E9 (R1) addendum on estimands [7]. With perfect adherence to the assigned treatment strategy and study protocol, results from ITT, PP, and AT analyses would be identical. Compared to results from PP and AT analyses, those from ITT analyses can theoretically provide unbiased estimates of the randomly assigned treatment regardless of the adherence [2], but they may increasingly deviate from results from PP and AT analyses with increasing nonadherence. 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 that 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 confounders that change over time, i.e., time-varying confounders, such as patient characteristics (e.g., body weight) or even the treatment that the study aims to explore [2,8]. This can, e.g., be a demanding workout intervention that some participants adhere to and some do not, e.g., because they are unsatisfied with its effect on weight. In such cases, standard approaches for confounder control could be inappropriate [9]. Marginal structural model (MSM) analyses are used to adjust for confounding in observational research [9,10], and they can address time-varying confounding. If the relevant confounders are known, measured, and adequately implemented in the modeling [9], MSMs should theoretically allow to provide valid estimates of PP effects and also ITT effects.

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 [11]. We specifically focused on the relationship of MSM-based results and results from ITT analyses.

2. Methods

2.1. Search

We conducted four separate searches. First, we searched PubMed using textwords related to MSMs (including "IPTW" or "inverse probability") and the medical subject heading for MSMs applying the Cochrane sensitivity- and precision-maximizing RCT filter [12] (Web Appendix 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

What is new?

Key findings

- MSMs typically provided largely 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.

What this adds to what was known?

- Despite conceptual differences, results from marginal structural modeling and intention-to-treat analyses often came to similar conclusions in clinical trials.
- The spread among numerous reported effect estimates for the same outcome, even within conventional analyses, can sometimes be substantial, and incremental benefits of complex analyses such as marginal structural models may be neutralized without maximal transparency and safeguards to avoid selective reporting bias

What is the implication and what should change now?

- Marginal structural modeling may provide helpful insights and different perspectives on treatment effects in the analysis of randomized controlled trials. Introduction of such complex methods require more detailed and strict measures as safeguards to avoid research-associated biases such as selective reporting. Otherwise, possible theoretical advantages may be entirely neutralized.

12 key references (selected by expert opinion of the authors group) in the field of MSMs [13−24]. Fourth, we also used and updated the search strategy from a related ongoing project in which we compared the effect estimates from nonrandomized studies using MSMs with those from systematically identified RCTs not using MSMs. On title−abstract level, we considered any study reporting on MSMs or using any form of inverse probability weighting as potentially relevant. All full-text publications were assessed by two independent reviewers (H.E., and either A.L. or B.S.), and disagreements were resolved by discussion or with a third reviewer (L.G.H.).

2.2. Selection of studies

We included any RCT (including reanalyses of RCTs) that reported the effect estimates of any health care intervention analyzed using MSMs and at least one effect estimate from an ITT, AT, or PP analysis. We relied on authors' definitions of analysis approaches (including specification of ITT, e.g., modified ITT or ITT for sensitivity analysis). When we were unsure whether a reported effect estimate was analyzed 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 if it was used to analyze the randomized treatment comparison (response rate 52%). For 23 effect estimates from 8 RCTs [25−32] where we could not clearly determine the ITT effect estimate, 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 and asked the main study authors to confirm or send us the protocol or information on its retrieval. The protocols were used to understand why MSM approaches were chosen, to obtain supplemental information on prespecification 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 (H.E. and B.S.) independently screened the reference lists of the MSM publications, trial homepages, PubMed, and Clinicaltrials.gov.

2.3. Data extraction

From each eligible RCT, we selected all clearly MSM-based effect estimates on any outcome using any metric (in one case [33], both risk difference and hazard ratio were reported and we extracted only the hazard ratio). For each reported MSM-based effect estimate, we identified any corresponding non-MSM-based effect estimate in the same publication and the main trial publication (where applicable) that was based on the same clinical question (i.e., population, intervention, control, and outcome) and follow-up timepoint (allowing for up to 12 months deviance). We specifically identified any effect estimate 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 [CIs]), and details on the analysis approaches. For two clinical questions of one trial with continuous outcomes, there was no between-group difference and we calculated it using the reported changes from baseline [34,35]. We extracted the effect estimates for the overall trial population where possible. In one case, we extracted the results for two mutually exclusive subpopulations (aspirin users and nonusers) as no MSM effect estimate was reported for the overall population [28]. In three other cases, the MSM-based effect estimate was only reported for a subpopulation of the main trial [27,36,37], and we only used non-MSM analyses for the same subpopulation.

We extracted general trial characteristics and determined the primary endpoint and whether an MSM analysis was prespecified 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.

2.4. Data analysis

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

First, using the results from all available analyses, we assessed how frequently treatment effect estimates 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 estimate lay within the 95% CI of the other effect estimates. We also determined the overall vibration of treatment effect estimates per clinical question, i.e., the spread between the largest and smallest effect sizes (on a relative risk [odds ratio or hazard ratio] scale) derived from different analytical methods on the same clinical question [38,39]. The vibration was determined excluding two trials that only had effect estimates for continuous outcomes.

Second, to specifically focus on MSM- vs. ITT-based results across all clinical questions, we selected the main MSM- and main ITT-based effect estimates for each clinical question. When multiple variations of such effect estimates were reported, we selected the one described as "main" or "primary" (in the MSM publication for the MSM-based effect estimate and in the main publication for the ITT effect estimate). 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).

Third, to specifically compare the MSM- and ITT-based results on a trial level, we selected one main clinical question of each trial. When there were multiple clinical questions 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 clinical questions (one trial compared two interventions with one control [40] and another used MSMs for two mutually exclusive subpopulations [28]).

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 effect estimates. We tested if one approach more frequently provided more extreme effect estimates than the other with the test for one proportion [41]. 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 differences from the null are identical and the ratio of the deviations is onefold. A ratio of >1 indicates more extreme effect estimates for MSM-based results.

Finally, we determined how similar the estimates of MSM vs. ITT analyses are using the ratio of the estimated relative risks (by calculating the absolute difference between MSM- and ITT-based effect sizes on the log-scale). For example, if the relative risk estimates are 0.5 and 2.0 with the two approaches, the ratio of the estimated relative risks is fourfold. This ratio is >1 by definition as it reflects the absolute difference between both estimates.

We also determined if MSM-based relative risk estimates were more or less precise than ITT-based effect estimates by calculating the ratio of standard errors of both approaches.

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.

2.5. Patient and public involvement

No patients/public were involved in this research.

3. Results

The search yielded 4372 records (last searched 19 May 2017), 176 were assessed in full text. We included 14

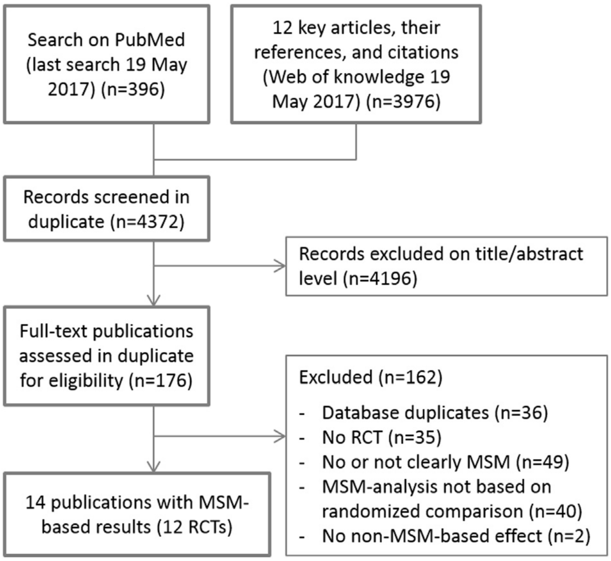
Search on PubMed (last search 19 May 2017) (n=396)
12 key articles, their references, and citations (Web of knowledge 19 May 2017) (n=3976)
Records screened in duplicate (n=4372)
Records excluded on title/abstract level (n=4196)
Full-text publications assessed in duplicate for eligibility (n=176)
Excluded (n=162)
- Database duplicates (n=36)
- No RCT (n=35)
- No or not clearly MSM (n=49)
- MSM-analysis not based on randomized comparison (n=40)
- No non-MSM-based effect (n=2)
14 publications with MSM-based results (12 RCTs)

Fig. 1. Study flow. *Abbreviations*: MSM, marginal structural models; RCT, randomized controlled trial.

Table 1. Characteristics of included studies

RCT	No. randomized	Patients' condition	Intervention and control	Outcomes with MSM-based results analytic approach (n)	Total number of pertinent clinical questions
ACTG 320[a] [29,42]	1156	HIV positive, immunosuppressed, ART-experienced patients	HAART (Zidovudine and Lamivudine plus Indinavir) vs. CART (Zidovudine and Lamivudine)	AIDS or death (primary) MSM (8) ITT (3)	11
ARISTOTLE [28,43,44]	18,201 (using aspirin at BL: 5632)	Atrial fibrillation (in aspirin users and nonusers)	Apixaban vs. Warfarin	Stroke or systemic embolism (primary, subgroups only) MSM (1) ITT (2) Major bleeding MSM (1) As treated (2)	6
	18,201 (not using aspirin at BL: 12,569)			Stroke or systemic embolism (primary, subgroups only) MSM (1) ITT (2) Major bleeding MSM (1) As treated (2)	6
CALERIE [34,35]	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)	RMR (primary) MSM (1) ITT (2) Core temperature (primary) MSM (1) ITT (2)	6
Kisumu[a] [25,26,45]	2784	Uncircumcised, HIV-negative young men	Immediate vs. delayed circumcision	HIV incidence (primary) MSM (1) As treated (2) Herpes simplex virus 2 incidence MSM (1) As treated (2)	6[b]
Negoro/ Yamaguchi [36,46][c]	398 (MSM analysis only for 2 of 3 groups with 266 patients)	Stage IIIB lung cancer (NSCLC)	Irinotecan hydrochloride vs. cisplatin	Overall survival (primary) MSM (1) ITT (3)	4
PHS[a] [30,32,47]	22,071	Male physicians	Aspirin vs. placebo	Cardiovascular mortality (primary) MSM (3) ITT (4) As treated (3)	10
PointBreak [48–50]	939	Stage IIIB or IV lung cancer (NSCLC)	''Pemetrexed/ Carboplatin/ Bevacizumab followed by maintenance Pemetrexed/ Bevacizumab'' vs. ''Paclitaxel/ Carboplatin/ Bevacizumab followed by Maintenance Bevacizumab''	Overall survival (primary) MSM (3) ITT (3)	6

(*Continued*)

Table 1. Continued

RCT	No. randomized	Patients' condition	Intervention and control	Outcomes with MSM-based results analytic approach (n)	Total number of pertinent clinical questions
PREDIMED[a] [27,40]	7447 (non-diabetic subgroup: 3833)	Risk factors for CVD	Mediterranean diet supplemented with extra-virgin olive oil vs. advice on a low-fat diet	Type 2 diabetes mellitus incidence MSM (1) ITT (11)	12
			Mediterranean diet supplemented with nuts vs. advice on a low-fat diet	Type 2 diabetes mellitus incidence MSM (1) ITT (11)	12
Ranapurwala [51]	1660 (70 randomized units)	Recreational scuba divers	Checklist vs. no checklist	Any diving mishap (primary) 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)	18
Tunis/Faries [37,52]	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	Change in brief psychiatric rating scale (primary) MSM (1) ITT (2) On drug (4) Epoch (2)	9
WHI[a] [33,53,54]	16,608	Postmenopausal women with intact uterus	Estrogen-plus-progestin vs. placebo	Coronary Heart Disease (primary) MSM (1) ITT (2) Invasive breast cancer incidence MSM (1) ITT (3)	7
WHS[a] [31,55,56]	39,876	Female health professionals	Aspirin vs. placebo	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)	25

(*Continued*)

Table 1. Continued

RCT	No. randomized	Patients' condition	Intervention and control	Outcomes with MSM-based results analytic approach (n)	Total number of pertinent clinical questions
				Cardiovascular disease mortality MSM (1) ITT (3) As treated (1)	

Abbreviations: 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.

[a] Trial stopped early.

[b] The main publication is based on 2-year follow-up (effect estimates not considered).

[c] 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 study compared 3 treatment arms of which 2 were different doses of Irinotecan. As MSM-based results were only available for the clinical question with the lower-dose Irinotecan (60 mg m^{-2}), we do not present the third arm (100 mg m^{-2}).

publications reporting results of 12 RCTs with a median of 1972 included patients (interquartile range [IQR] 870 to 17,006) (Fig. 1; Table 1). They were published between 2002 and 2016 (median 2013). Six of the 12 RCTs stopped early, 4 for benefit [32,40,42,45] and 2 for harm [31,53]. The studies evaluated treatment effect estimates 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 [31,32,53], no intervention [34,51], or delayed intervention [45]. They reported outcomes related to cardiology [31,32,43,53], oncology [46,48,53], infectious diseases [29,42,45], diabetes [27], psychiatry [52], gerontology [34], and physical education [51] (Table 1). Double blinding was reported in 6 of 12 RCTs [31,32,42,43,45,53].

For 8 RCTs, we identified a protocol [31,32,34,40,43,53] or design paper [42], and only in one of them, we found a clear prespecification of MSMs [34] (Web Appendix 2). 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 [42,48,53]. The MSM publication was published a median of 3 years after the main trial publication. The stated motivations for applying MSMs were diverse: MSMs were used to adjust for "time-dependent" or "time-varying" confounding [26,28−30,33,35−37,49,55,57], "noncompliance" or "nonadherence" [28,29,33,35,51,55,57], "loss to follow-up" [26], treatment switching [37], second-line treatment [36], and "to analyze the data as if it were from an observational study rather than a randomized, controlled trial" [27] (Web Appendix 2). All RCTs reported fitting a form of Cox or logistic model, 4 reported the use of several models; 11 of the 12 RCTs reported inverse probability weighting, the other RCT [28] reported only "weighting" in relation to MSMs (further details in Web Appendix 3).

Across the 12 RCTs, we identified 24 clinical questions (median 6; IQR 3 to 7 per question). Overall, 138 analyses were reported for these 24 questions of which 38 were MSM based (including sensitivity analyses, "crude" and adjusted analyses, different censoring, and different forms of MSMs). 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) (Fig. 2). Twenty-one clinical questions had binary outcomes, and 3 clinical questions (2 RCTs) had continuous outcomes.

Two analyses using MSMs were clearly prespecified (1 trial), and 4 analyses using MSMs were explicitly described as "sensitivity analysis" (2 trials). MSMs were used to evaluate the primary endpoint in 11 of the 12 RCTs.

3.1. Overall relationship of treatment effect estimates

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

Among the 123 analyses reported for 21 clinical questions with binary outcomes, the median spread between the largest and smallest effect estimates 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.

3.2. Relationship of MSM- and ITT-based results

MSM- and ITT-based results were all in the same direction across all 20 available clinical questions (100%; Table 2). Their CIs overlapped in all 18 cases with available CI

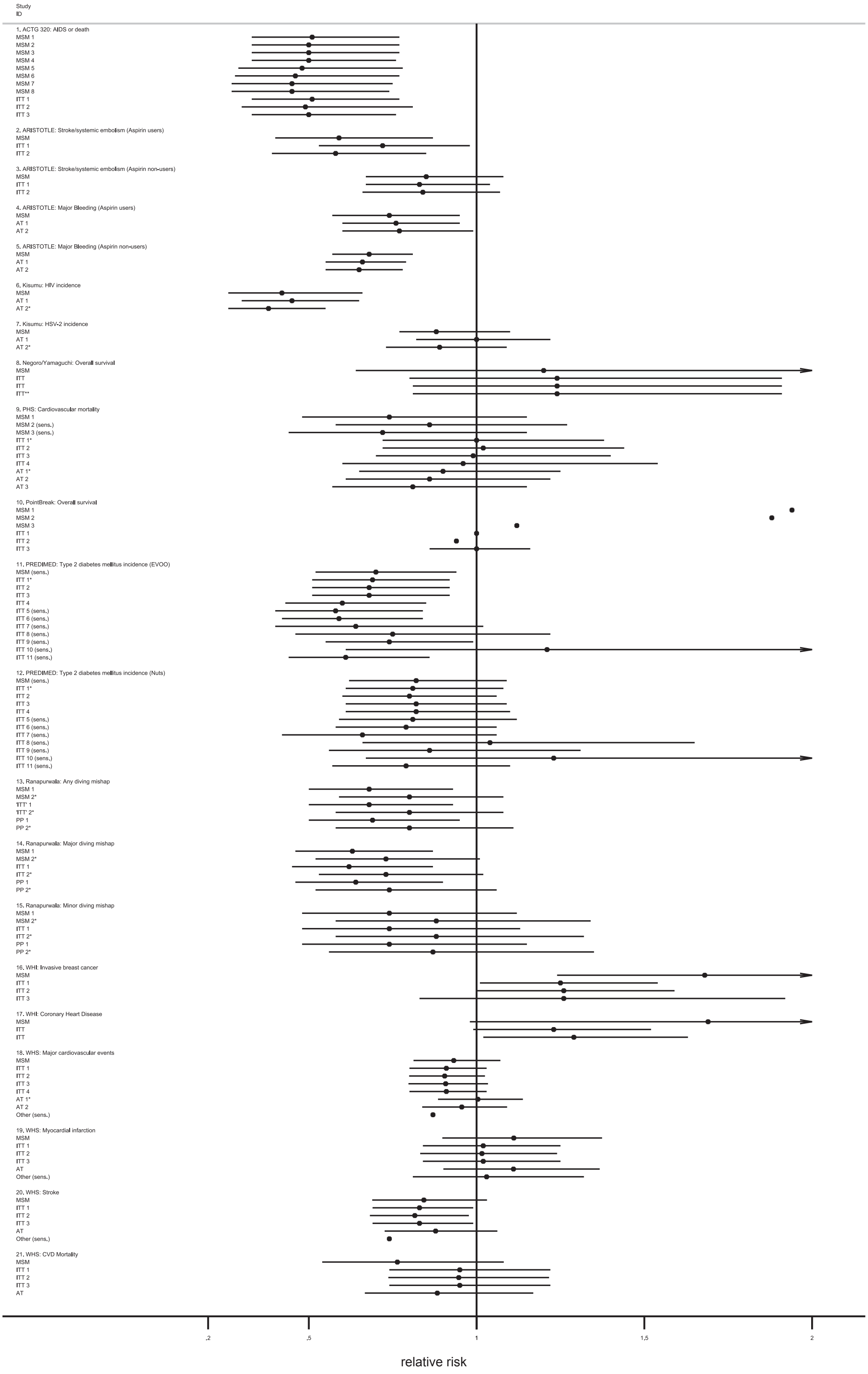

Fig. 2. Overview of results from different analyses for the same clinical question (population, intervention, control, outcome) for all 24 clinical questions reported in the main publication and the publication with MSM results. Circles indicate effect estimates, and lines 95% confidence intervals (in some cases not reported). *Abbreviations*: 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; * unadjusted analyses.

Table 2. Relationship of effect estimates per outcome

Clinical question	Vibration of treatment effect estimates: Spread of lowest vs. highest relative risk estimate across all reported analyses[a]	Ratio of deviations from the null with MSM and ITT (x-fold more extreme effect estimates with MSM)[a,e]	Ratio of the relative risks with MSM and ITT[a] (x-fold difference between effect estimates)
ACTG 320: AIDS or death	1.13	1.11	1.11
ARISTOTLE (Aspirin nonusers): Major Bleeding	1.05	NA	NA
ARISTOTLE (Aspirin users): Major Bleeding	1.04	NA	NA
ARISTOTLE (Aspirin nonusers): Stroke or systemic embolism	1.02	0.98	1.02
ARISTOTLE (Aspirin users): Stroke or systemic embolism	1.24	1.22	1.22
Kisumu: HIV incidence	1.18	NA	NA
Kisumu: HSV-2 incidence	1.14	NA	NA
Negoro/Yamaguchi: Survival	1.03	0.97	1.03
PREDIMED (EVOO): Incidence type 2 diabetes mellitus	2.09	0.99	1.01
PREDIMED (Nuts): Incidence type 2 diabetes mellitus	1.86	0.99	1.01
PHS: CVD mortality	1.42	1.3	1.3
PointBreak: Overall survival	2.06	1.12	1.12
Ranapurwala: All mishaps	1.18	1.18	1.18
Ranapurwala: Major mishaps	1.19	1.16	1.16
Ranapurwala: Minor mishaps	1.19	1.19	1.19
WHI: Coronary heart disease	1.37	1.31	1.31
WHI: Invasive breast cancer	1.34	1.33	1.33
WHS: CVD mortality	1.24	1.24	1.24
WHS: Major CVD events	1.15	0.98	1.02
WHS: Myocardial infarction	1.09	1.09	1.09
WHS: Stroke	1.19	0.98	1.02
Tunis/Faries: change in BPRS	NA	NA	NA
CALERIE: Resting metabolic rate	NA	NA	NA
CALERIE: Core Temperature	NA	NA	NA
Median (IQR) or total (%)	1.19 (1.13 to 1.34)[d]	1.12 (0.99 to 1.22)	1.12 (1.02 to 1.22)
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)

Abbreviations: 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.

[a] Dichotomous outcomes only.

[b] No 95% CI for the MSM-based result reported.

[c] No 95% CI for the MSM-based nor the ITT-based result reported.

[d] The median (IQR) excluding sensitivity analyses is 1.17 (1.08 to 1.24).

[e] >1 indicates more extreme effect estimates for MSM-based results.

information (100%), and the MSM effect estimate lay within the 95% CI of ITT effect estimates in 16 cases (89%). Twelve of 17 (71%, 3 cases with at least 1 CI missing) had both the same direction of effect estimate and were both nominally significant or both nominally nonsignificant (i.e., both 95% CIs included the null or not; Table 2).

MSM-based effect estimates were more extreme in 13 of 20 clinical questions (65%) and in 7 of 20 (35%), and ITT-based effect estimates were more extreme ($P = 0.18$). The median deviation from the null of the MSM-based effect estimates was 1.35 (IQR 1.19 to 1.59) and of ITT effect estimates 1.24 (IQR 1.10 to 1.29) on a relative risk scale. On

MSM and ITT effect estimates in same direction	MSM and ITT effect estimates with same stat. significance	MSM and ITT effect estimate CI overlapping	MSM effect estimate within CI of ITT effect estimate
Yes	Yes	Yes	Yes
NA (no ITT effect estimate)	NA (no ITT effect estimate)	NA (no ITT effectestimate)	NA (no ITT effect estimate)
NA (no ITT effect estimate)	NA (no ITT effect estimate)	NA (no ITT effect estimate)	NA (no ITT effect estimate)
Yes	Yes	Yes	Yes
Yes	Yes	Yes	Yes
NA (no ITT effect estimate)	NA (no ITT effect estimate)	NA (no ITT effect estimate)	NA (no ITT effect estimate)
NA (no ITT effect estimate)	NA (no ITT effect estimate)	NA (no ITT effect estimate)	NA (no ITT effect estimate)
Yes	Yes	Yes	Yes
Yes	Yes	Yes	Yes
Yes	Yes	Yes	Yes
Yes	Yes	Yes	Yes
Yes	NA[b]	NA[b]	Yes
Yes	No	Yes	Yes
Yes	No	Yes	Yes
Yes	Yes	Yes	Yes
Yes	No	Yes	No
Yes	Yes	Yes	No
Yes	Yes	Yes	Yes
Yes	Yes	Yes	Yes
Yes	Yes	Yes	Yes
Yes	No	Yes	Yes
Yes	No	Yes	Yes
Yes	NA[c]	NA[c]	NA[c]
Yes	NA[c]	NA[c]	NA[c]
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%)
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%)

average (median), the ratio of these deviations indicated 1.12-fold more extreme MSM-based effect estimates than the corresponding ITT effect estimates (IQR 0.99 to 1.22; Table 2).

When analyzing only the 13 main clinical questions, MSM-based effect estimates were more extreme than ITT-based effect estimates in 7 questions (54%). In 46%, ITT-based effect estimates were more extreme ($P = 0.78$). The median deviation from the null of the MSM-based effect estimates was 1.39 (IQR 1.19 to 1.69) and of ITT effect estimates 1.24 (IQR 1.15 to 1.34). Here, MSM-based effect estimates were 1.11-fold more extreme (IQR 0.98 to 1.20; Table 2).

The absolute 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 effect estimates deviated at least 1.12-fold from ITT effect estimates. Among the 11 main clinical questions, this was 1.11-fold (IQR 1.02 to 1.20; Table 2). Details of original study effect estimates from main MSM- and ITT-based analyses of the 12 trials are in Web Appendix 4.

The precision of effect estimates from MSM and ITT was very similar (median ratio of standard errors 1.01-fold; IQR 1.00 to 1.04).

4. Discussion

4.1. Principal findings

In this empirical analysis, we found 12 trials with 138 effect estimates for 24 clinical questions that reported results from MSM-based and conventional analyses (Fig. 2). The main motivations for using MSMs were related to time-varying confounding including nonadherence. The differences between MSM-based and other effect estimates, including ITT effect estimates, were typically within chance, and the effect estimates were in the same direction. However, the quantitative differences across reported effect estimates even within the same trial for the same outcome and the same author groups using different methods can be substantial sometimes. The MSM used in the context of these trials does not consistently yield more extreme effect estimates than ITT. Overall, MSM and ITT effect estimates were similar, and the absolute difference was less than 1.12-fold in half of the clinical questions. However, although 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 in the context of these trials 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 clinical questions of treatment effect estimates), it may make a substantial difference on which analysis method is chosen. When it comes to weighing benefits and harms of treatments or informing shared decisions, e.g., 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 health care decision-makers to know which to base their decisions on. In one study, e.g., there were 11 ITT effect estimates, and many studies had two or more ITT effect estimates. The analyses for these estimates followed different approaches and had different degrees of statistical adjustments (e.g., crude and adjusted for various covariates). For the outcomes used here, we found for only 1 study (CALERIE) a clear prespecification of the use of MSMs. The authors clearly report their motivation to use MSMs and state that they aimed to address the mechanistic question "What is the direct physiological effect of calorie restriction?" [58]. Answering this kind of mechanistic question (and clearly labeling it as such) is a very insightful addition and may help to explore the theoretical impact of the treatment under perfect conditions and to generate new research hypotheses. This illustrates the potential value of using MSMs for trials in different situations, as an MSM-based PP effect estimate would be unbiased, albeit under the strong assumptions that all confounders are known, measured, and implemented correctly in the model.

For only 5 studies, we found a clear statement on the use of ITT in their protocol or design publications. This could add uncertainty to the interpretation of trial results [59,60]. In such a setting, the substantial vibration of effect estimates can offer opportunities for sizable 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 prespecification 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 timepoint as prespecified in protocols and trial registries, the results from various statistical approaches may provide different effect sizes and can be selectively reported.

For one study, PREDIMED [40], we recently learned that there may have been problems with the randomization and that the main publication was retracted and corrected [61]. The outcome we used from this study was, however, not reported in this publication or its correction. Hence, we do not know how our results would have been influenced by this. We conducted a sensitivity analysis in which we excluded PREDIMED and found no relevant changes on our main results (data not shown).

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 MSMs was often an explorative approach. However, conducting several analytical methods in addition to the prespecified analyses, especially methods as complex as MSMs 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 prespecified, statistical details of applying such a complex approach may still affect the results. Various quality control procedure measures have been proposed to prevent such selective outcome reporting [60].

4.2. Comparison with other studies

This is, as far as we know, the first meta-epidemiological analysis comparing the results from causal modeling 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 [62−64]. A reanalysis of an RCT evaluating interventions for symptom management compared the conclusions from ITT (without imputing missing data) with those from PP analysis [62]. While the conclusions did not differ, the PP analysis also indicated which intervention and dose strategies affected symptoms [62]. A systematic review of RCTs reporting both ITT and PP analyses on a primary binary endpoint found effect estimates from PP analyses to be more extreme and the ratio to ITT analyses varied greatly (0.39 to 2.53) [63]. 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 estimate of either ITT or PP alone [63]. 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 effect estimates [65]. An analysis of 200 randomized trials published in the high-impact-factor journals in 2009 showed that primary outcomes are often analyzed in different ways and that the nominal statistical significance would change in about one of five studies (18%), depending on the adjustment for stratification variables and baseline characteristics [59].

4.3. Limitations

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

Second, we encountered various different forms and descriptions of MSMs, e.g., standard MSM [20], augmented MSM, adaptively truncated MSM [49], MSM for binary and continuous outcomes [37], using IPTW, IPCW, IPW, G-estimation, adjusted or "unadjusted" [51], censored at different time points [29], and adjusted for different covariates with or without two-way interactions between randomization status and each covariate [29]. Some of our included studies even reported the results of multiple different forms of MSMs within their study [29,30,49,51]. We also aimed to obtain some details on the weighting, but overall we were not able to explore the agreement between effect sizes in relation to all these factors.

Third, we did not verify the analytic approaches and relied on the authors' descriptions of them. We also did not try to assess the validity or quality of the methods applied. There is to our knowledge no tool to allow an assessment of the validity of the analyses. Also, the reporting is widely inconsistent, which makes proper classification from our side very difficult (see below). Although we believe that the authors are probably the best experts for their data and analyses and have correctly applied the methods, classified, and described them, details of the definitions may be inconsistent between different reports [66].

Fourth, the trials that applied MSMs to analyze the randomized treatment 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 [34,46,51] 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 [52]). Many (5/12) trials were also discontinued early, more frequently than in the typical clinical trial literature [67]. Hence, our sample appears not to be representative of all RCTs.

Fifth, we encountered problems with vague reporting of the analysis methods used. For example, 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 [66].

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. Frequent inconsistent reporting of ITT approaches is well described and may also vary across medical fields [66,68,69]. There is also no consensus on issues of missing data related to ITT analyses—however, none of the study authors reported "modified ITT" [65] analyses in our sample. As MSM-IPTW methodology becomes more widely used and users may not be as experienced, the quality of the methods used and choices such as the mean and range of weight estimates and handling of extreme weights may become important in shaping the exact results. We were not able to explore the agreement between effect sizes in relation to these factors.

Seventh, the application of MSMs in many studies was for very different reasons. MSMs were 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 health care decision-making. This further adds to the very limited generalizability of this small but, nevertheless, systematically derived sample of trials.

Eighth, there were only four trials reporting (non-MSM-based) results from AT analyses, and only one trial reporting (non-MSM-based) results from PP analyses. We would need more data to explore specific differences between MSM-based estimates and the results from AT and PP analyses.

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

MSM analyses require more sophisticated modeling than ITT analyses. It is difficult to prespecify and collect the detailed high-quality data that are required for analyzing all possible postrandomization confounders, such as nonadherence [70]. Because patients' preferences and values leading to nonadherence are almost never included in data collection, important confounders very likely remain unmeasured and hence cannot be included in the modeling. Therefore, probably, there is always some residual confounding bias even in MSM-adjusted effect estimates. Furthermore, caution is required to prespecify 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 do not require such adjustments and are more straightforward to prespecify. 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 current "real-world" research conditions.

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 on treatment effects, especially when there are high rates of attrition and nonadherence. However, there is a wide spread across all reported effect estimates 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 and helpful responses to our information requests. We also thank Soheila Aghlmandi (University of Basel) for providing external feedback on statistical unclarities and Benjamin Kasenda (Department of Oncology, University Hospital Basel) for providing subject matter expertise for oncology topics.

Authors' contributions: H.E. and L.G.H. conceived the study with input on the study design by J.P.A.I. and H.C.B. H.E. conducted the literature searches. H.E., L.G.H., B.S., and A.L. extracted the data. H.E., L.G.H., and J.P.A.I. analyzed the data. H.E., L.G.H., and J.P.A.I. interpreted the results. H.E. wrote the first draft, and all authors made revisions on the manuscript. All authors read and approved the final version of the article. H.E. acquired funding for this study. H.E. and L.G.H. are the guarantors.

Supplementary data

Supplementary data related to this article can be found at https://doi.org/10.1016/j.jclinepi.2018.11.001.

References

[1] Hollis S, Campbell F. What is meant by intention to treat analysis? Survey of published randomised controlled trials. BMJ 1999; 319(7211):670–4.

[2] Hernán MA, Hernández-Díaz S. Beyond the intention to treat in comparative effectiveness research. Clin Trials 2012;9(1):48–55.

[3] Senn S. Seven myths of randomisation in clinical trials. Stat Med 2013;32:1439–50.

[4] Karanicolas PJ, Montori VM, Devereaux PJ, Schünemann H, Guyatt GH, et al. A new 'Mechanistic-Practical' Framework for designing and interpreting randomized trials. J Clin Epidemiol 2009;62:479–84.

[5] Ioannidis JPA. Randomized controlled trials: often flawed, mostly useless, clearly indispensable: a commentary on Deaton and Cartwright. Soc Sci Med 2018;210:53–6.

[6] Hernán MA, Robins JM. Causal Inference, Part I, [chapter 9]. 5 Per-protocol effect 2018:117. Available at https://www.hsph.harvard.edu/miguel-hernan/causal-inference-book/. Accessed August 30, 2018.

[7] 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. Available at http://www.ema.europa.eu/ema/doc_index.jsp?curl=pages/includes/document/document_detail.jsp?webContentId=WC500233916&murl=menus/document_library/document_library.jsp&mid=0b01ac058009a3dc. Accessed August 30, 2018.

[8] Mansournia MA, Etminan M, Danaei G, Kaufman JS, Collins G. Handling time varying confounding in observational research. BMJ 2017;359:j4587.

[9] Williamson T, Ravani P. Marginal structural models in clinical research: when and how to use them? Nephrol Dial Transplant 2017;32(suppl_2):ii84–90.

[10] 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):1–6.

[11] Hemkens LG, Contopoulos-Ioannidis DG, Ioannidis JP. Routinely collected data and comparative effectiveness evidence: promises and limitations. CMAJ 2016;188(8):E158–64.

[12] Lefebvre C, Manheimer E, Glanville J. 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. Available at www.handbook.cochrane.org.

[13] Cole SR, Hernan MA. Constructing inverse probability weights for marginal structural models. Am J Epidemiol 2008;168:656–64.

[14] 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.

[15] Hernán MA, Brumback B, Robins JM. Marginal structural models to estimate the joint causal effect of nonrandomized treatments. J Am Stat Assoc 2001;96:440–8.

[16] Hernan MA, Robins JM. Estimating causal effects from epidemiological data. J Epidemiol Commun Health 2006;60:578–86.

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

[18] Robins JM. Correction for non-compliance in equivalence trials. Stat Med 1998;17:269–302. discussion 87-89.

[19] Robins JM. Association, causation, and marginal structural models. Synthese 1999;121(1–2):151–79.

[20] 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; 2000:95–133.

[21] 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. J Am Stat Assoc 1999;94:687–700.

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

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

[25] 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. J Infect Dis 2013;208:1869–76.

[26] 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.

[27] 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. Ann Intern Med 2014;160:1.

[28] 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.

[29] 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.

[30] 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.

[31] 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.

[32] 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.

[33] 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.

[34] 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(9):1097–104.

[35] 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.

[36] 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.

[37] Faries D, Ascher-Svanum H, Belger M. Analysis of treatment effectiveness in longitudinal observational data. J Biopharm Stat 2007;17(5):809–26.

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

[39] Patel CJ, Burford B, Ioannidis JPA. Assessment of vibration of effects due to model specification can demonstrate the instability of observational associations. J Clin Epidemiol 2015;68:1046–58.

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

[41] Social Science Computing Cooperative. Stata for Students: Proportion Tests 2016. Available at https://www.ssc.wisc.edu/sscc/pubs/sfs/sfs-prtest.htm. Accessed August 30, 2018.

[42] 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.

[43] 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.

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

[45] 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.

[46] 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.

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

[48] 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.

[49] 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(6):448–54.

[50] 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(4):252–6.

[51] 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 2016;45:223–31.

[52] 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 Health 2006;9(2):77−89.

[53] 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.

[54] 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.

[55] 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(6):431−8.

[56] Buring JE, Hennekens CH. The Women's health study: summary of the study design. J Myocardial Ischemia 1992;4:27−9.

[57] 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.

[58] Protocol to CALERIE study. Available at https://calerie.duke.edu/files/phase2_protocol.pdf. Accessed August 30, 2018.

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

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

[61] Estruch R, Ros E, Salas-Salvadó J, Covas M-I, Corella D, Arós F, et al. Retraction and republication: primary prevention of cardiovascular disease with a Mediterranean diet. N Engl J Med 2013;368:1279-1290. N Engl J Med 2018;378:2441−2.

[62] 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(6):E293−302.

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

[64] 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.

[65] 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.

[66] 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.

[67] 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.

[68] Bell ML, Fiero M, Horton NJ, Hsu CH. Handling missing data in RCTs; a review of the top medical journals. BMC Med Res Methodol 2014;14:118.

[69] Del Re AC, Maisel NC, Blodgett JC, Finney JW. Intention-to-treat analyses and missing data approaches in pharmacotherapy trials for alcohol use disorders. BMJ Open 2013;3(11):e003464.

[70] Hernan MA, Robins JM. Per-protocol analyses of pragmatic trials. N Engl J Med 2017;377:1391−8.