

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

---

Hannah Ewald<sup>a,c</sup>, Benjamin Speich<sup>a</sup>, Aviv Ladanie<sup>a,b</sup>, Heiner C Bucher<sup>a</sup>, John PA Ioannidis<sup>d,h</sup>, Lars G Hemkens<sup>a</sup>

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

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

*c University Medical Library, University of Basel, 4051 Basel, 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*

**Correspondence to:** [Lars.Hemkens@usb.ch](mailto:Lars.Hemkens@usb.ch), Lars G Hemkens, Basel Institute for Clinical Epidemiology & Biostatistics, University Hospital Basel, Spitalstrasse 12, CH-4031 Basel, Phone +41 61 32 85404, Fax +41 61 26 53109

## Abstract

### Objective

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.

### 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% 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; IQR 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; IQR 1.02-1.22).

### 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

**Running title:** Marginal structural models in randomized clinical trials

**Count:** Abstract: 227, Text: 4958

Number of figures: 2

Number of tables: 2

Number of webappendices: 4

## 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 is known

- Despite conceptual differences, results from marginal structural modelling and intention-to-treat analyses often came to largely 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, what should change now

- Marginal structural modelling 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.

## 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 <sup>1</sup>, 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 <sup>2</sup>.

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 trial protocol. Randomization makes confounding random and this hopefully improves the chances of getting valid estimates of effects of such decisions<sup>3-5</sup>. Such trials focusing on health care decision making are often pragmatic or practical <sup>4</sup>. 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<sup>2</sup>) or AT analyses (evaluating patients according to the treatment they received, not the treatment they were assigned to<sup>2</sup>) 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.”<sup>6</sup>.

The conceptual difference of 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 <sup>7</sup>. 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, ITT analyses can theoretically provide unbiased estimates of the randomly assigned treatment regardless of the adherence <sup>2</sup>, but they may increasingly deviate from results from PP and AT analyses 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 confounders that change over time, i.e. time-varying confounders, such as patient characteristics (for example body weight) or even the treatment that the study aims to explore <sup>2 8</sup>. This can, for

example, 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<sup>9</sup>. Marginal structural model (MSM) analyses are used to adjust for confounding in observational research<sup>9,10</sup> and they can address time-varying confounding. If the relevant confounders are known, measured and adequately implemented in the modelling<sup>9</sup>, MSM 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<sup>11</sup>. We specifically focused on the relationship of MSM-based results and results from ITT analyses.

## 2. Methods

### Search

We conducted four separate searches. First, we searched PubMed using textwords related to MSM (including "IPTW" or "inverse probability") and the medical subject heading for MSM applying the Cochrane sensitivity- and precision-maximizing RCT filter<sup>12</sup> (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<sup>13-24</sup>. Fourth, we also used and updated the search strategy from a related ongoing project in which we compared the effect estimates from non-randomized studies using MSMs with those from systematically identified RCTs not using MSMs. On title-abstract level, we considered any study reporting on MSM or using any form of inverse probability weighting as potentially relevant. 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 effect estimates of any health care intervention analyzed using MSM and at least one effect estimate from an ITT, as treated or per protocol 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<sup>25-32</sup> 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 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 effect estimates on any outcome using any metric (in one case<sup>33</sup>, 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, outcome) and follow-up time-point (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), 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<sup>34 35</sup>. 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 non-users) as no MSM-effect estimate was reported for the overall population<sup>28</sup>. In three other cases, the MSM-based effect estimate was only reported for a subpopulation of the main trial<sup>27 36 37</sup> 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 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 size (on a relative risk [odds ratio or hazard ratio] scale) derived from different analytical methods on the same clinical question<sup>38 39</sup>. The vibration was determined excluding two trials which only had effect estimates for continuous outcomes.

Secondly, to specifically focus on MSM-based versus ITT-based results across all clinical questions, we selected the main MSM- and main ITT-based effect estimate 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).

Thirdly, 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<sup>40</sup> and another used MSM for two mutually exclusive subpopulations<sup>28</sup>).

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<sup>41</sup>. 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 effect estimates 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 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.

## 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 publications reporting results of 12 RCTs with a median of 1972 included patients; IQR 870 to 17006) (Figure 1; Tables

Table 1). They were published between 2002 and 2016 (median 2013). Six of the 12 RCTs stopped early, 4 for benefit<sup>32 40 42 43</sup> and 2 for harm<sup>31 44</sup>. 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<sup>31 32 44</sup>, no intervention<sup>34 45</sup>, or delayed intervention<sup>42</sup>. They reported outcomes related to cardiology<sup>31 32 44 46</sup>, oncology<sup>44 47 48</sup>, infectious diseases<sup>29 42 43</sup>, diabetes<sup>27</sup>, psychiatry<sup>49</sup>, gerontology<sup>34</sup>, and physical education<sup>45</sup> (Table 1). Double blinding was reported in 6 of 12 RCTs<sup>31 32 42-44 46</sup>.

For 8 RCTs, we identified a protocol<sup>31 32 34 40 44 46</sup> or design paper<sup>43</sup>, and only in one of them we found a clear pre-specification of MSM<sup>34</sup> (Webappendix 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<sup>43 44 48</sup>. 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<sup>26 28-30 33 35-37 50-52</sup>, “non-compliance” or “non-adherence”<sup>28 29 33 35 45 50 52</sup>, “loss to follow-up”<sup>26</sup>, treatment switching<sup>37</sup>, second-line treatment<sup>36</sup>, and “to analyze the data as if it



were from an observational study rather than a randomized, controlled trial”<sup>27</sup>; Webappendix 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<sup>28</sup> reported only “weighting” in relation to MSM (further details in Webappendix 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 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 clinical questions had binary outcomes and 3 clinical questions (2 RCTs) had continuous outcomes.

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

### 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 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 clinical questions (100%; Table 2). Their CIs overlapped in all 18 cases with available CI 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 non-significant (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%), 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 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-based and ITT-based analyses of the 12 trials are in Webappendix 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

### Principal findings

In this empirical analysis, we found 12 trials with 138 effect estimates for 24 clinical questions which reported results from MSM-based and conventional analyses (Figure 2). The main motivations for using MSM were related to time-varying confounding including non-adherence. 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. 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, the absolute difference was less than 1.12-fold in half of the clinical questions. 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 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 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 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 pre-specification of the use of MSM. The authors clearly report their motivation to use MSM state that they aimed to address the mechanistic question “what is the direct physiological effect of calorie restriction?”<sup>53</sup>. Answering this kind of mechanistic question (and clearly labelling 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 MSM for trials in different situations, as an MSM-based per protocol 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<sup>54 55</sup>. In such a setting, the substantial vibration of effect estimates can offer opportunities for sizeable 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.

For one study, PREDIMED<sup>40</sup>, we recently learnt that there may have been problems with the randomization and that the main publication was retracted and corrected<sup>56</sup>. The outcome we used from this study was, however, not reported in this publication or its correction. So 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 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 affect the results. Various quality control procedures measures have been proposed to prevent such selective outcome reporting <sup>57</sup>.

### 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 <sup>58-60</sup>. 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 <sup>58</sup>. While the conclusions did not differ, the PP analysis also indicated which intervention and dose strategies affected symptoms <sup>58</sup>. A systematic review of RCTs reporting both ITT and PP analyses on a primary binary endpoint found effect estimates from PP analyses more extreme and the ratio to ITT analyses varied greatly (0.39 to 2.53)<sup>59</sup>. 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 <sup>59</sup>. 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 <sup>61</sup>. 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 <sup>54</sup>.

### Limitations

Our study has several limitations. First, we only identified 12 trials for which we had MSM- and non-MSM-based effect estimates and that focused on clinical decision making. Many of the excluded RCTs did not use MSM to analyze the randomized treatment 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 <sup>20</sup>, augmented MSM, adaptively truncated MSM <sup>51</sup>, MSM for binary and continuous outcomes <sup>37</sup>, using IPTW, IPCW, IPW, G-estimation, adjusted or “unadjusted” <sup>45</sup>, censored at different time points <sup>29</sup> and

adjusted for different covariates with or without two-way interactions between randomization status and each covariate <sup>29</sup>. Some of our included studies even reported the results of multiple different forms of MSM within their study <sup>29 30 45 51</sup>. 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 <sup>62</sup>.

Fourth, the trials that applied MSM 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 <sup>34 45 47</sup> 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 <sup>49</sup>). Many (5/12) trials were also discontinued early, more frequently than in the typical clinical trial literature <sup>63</sup>. 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 <sup>62</sup>.

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<sup>62 64 65</sup>. There is also no consensus on issues of missing data related to ITT analyses – however, none of the study authors reported "modified ITT"<sup>61</sup> 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 weights 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 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.

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 MSM or other specific models. While we found statements that clearly indicated such issues, the reporting quality was very heterogeneous.

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<sup>66</sup>. 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 effect estimates. 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 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 non-adherence. 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.

## Data sharing

No additional data available.

## Declaration of interests

All authors have completed the Unified Competing Interest form at [www.icmje.org/coi\\_disclosure.pdf](http://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 three years and no other relationships or activities that could appear to have influenced the submitted work.

## Contributors

HE and LGH conceived the study with input on the study design by JPAI and HCB. HE conducted the literature searches. 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. Benjamin Speich is supported by the Research Foundation of the University of Basel.

## 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.

## Copyright

The Corresponding Author has the right to grant on behalf of all authors and does grant on behalf of all authors, an exclusive licence (or non exclusive for government employees) on a worldwide basis to the BMJ Publishing Group Ltd to permit this article (if accepted) to be published in BMJ editions

and any other BMJPGL products and sublicences such use and exploit all subsidiary rights, as set out in our licence.

#### **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.



## 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**(7211):670-74.
2. Hernán MA, Hernández-Díaz S. Beyond the intention to treat in comparative effectiveness research. *Clinical trials (London, England)* 2012;**9**(1):48-55.
3. Senn S. Seven myths of randomisation in clinical trials. *Stat Med* 2013;**32**(9):1439-50.
4. Karanickolas PJ, Montori VM, Devereaux PJ, et al. A new 'Mechanistic-Practical' Framework for designing and interpreting randomized trials. *Journal of Clinical Epidemiology* 2009;**62**(5):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**:53-56.
6. Hernán MA, Robins JM. Causal Inference, Part I, chapter 9.5 Per-protocol effect, p.117. <https://www.hsph.harvard.edu/miguel-hernan/causal-inference-book/>. 2018. (accessed 30 August 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. [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](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 30 August 2018)
8. Mansournia MA, Etminan M, Danaei G, et al. 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? *Nephrology, dialysis, transplantation : official publication of the European Dialysis and Transplant Association - European Renal Association* 2017;**32**(suppl\_2):ii84-ii90.
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, 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.
13. Cole SR, Hernan MA. Constructing inverse probability weights for marginal structural models. *American journal of epidemiology* 2008;**168**(6):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. *Journal of the American Statistical Association* 2001;**96**(454):440-48.
16. Hernan MA, Robins JM. Estimating causal effects from epidemiological data. *J Epidemiol Community Health* 2006;**60**(7):578-86.
17. Robins J. Marginal structural models. 1997 Proceedings of the Section on Bayesian Statistical Science. Alexandria: American Statistical Association, 1998;1-10.
18. Robins JM. Correction for non-compliance in equivalence trials. *Stat Med* 1998;**17**(3):269-302; discussion 87-9.
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, eds. *Statistical Models in Epidemiology, the Environment, and Clinical Trials*. New York, NY: Springer New York, 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. *Journal of the American Statistical Association* 1999;**94**(447):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, Borrás 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, et al. Medical Male Circumcision and Herpes Simplex Virus 2 Acquisition: Posttrial Surveillance in Kisumu, Kenya. *Journal of Infectious Diseases* 2013;**208**(11):1869-76.
26. Mehta SD, Moses S, Agot K, et al. The long-term efficacy of medical male circumcision against HIV acquisition. *Aids* 2013;**27**(18):2899-907.
27. Salas-Salvado J, Bullo M, Estruch R, et al. Prevention of Diabetes With Mediterranean Diets A Subgroup Analysis of a Randomized Trial. *Annals of Internal Medicine* 2014;**160**(1):1.
28. Alexander JH, Lopes RD, Thomas L, et al. Apixaban vs. warfarin with concomitant aspirin in patients with atrial fibrillation: insights from the ARISTOTLE trial. *Eur Heart J* 2014;**35**(4):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**(12):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**(11):1045-53.
31. Ridker PM, Cook NR, Lee IM, et al. A randomized trial of low-dose aspirin in the primary prevention of cardiovascular disease in women. *N Engl J Med* 2005;**352**(13):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**(3):129-35.
33. Toh S, Hernandez-Diaz S, Logan R, et al. 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**(4):211-7.
34. Ravussin E, Redman LM, Rochon J, 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, et al. Application of the Marginal Structural Model to Account for Suboptimal Adherence in a Randomized Controlled Trial. *Contemp Clin Trials Commun* 2016;**4**:222-28.
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**(13):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**(5):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. *Journal of clinical epidemiology* 2015;**68**(9):1046-58.
40. Estruch R, Ros E, Salas-Salvado J, et al. Primary prevention of cardiovascular disease with a Mediterranean diet. *N Engl J Med* 2013;**368**(14):1279-90.
41. Social Science Computing Cooperative. Stata for Students: Proportion Tests. 2016. <https://www.ssc.wisc.edu/sscc/pubs/sfs/sfs-prtest.htm>. (accessed 30 August 2018)

42. Bailey RC, Moses S, Parker CB, et al. Male circumcision for HIV prevention in young men in Kisumu, Kenya: a randomised controlled trial. *Lancet* 2007;**369**(9562):643-56.
43. Hammer SM, Squires KE, Hughes MD, 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**(11):725-33.
44. Rossouw JE, Anderson GL, Prentice RL, 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**(3):321-33.
45. Ranapurwala SI, Denoble PJ, Poole C, et al. 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.
46. Granger CB, Alexander JH, McMurray JJ, et al. Apixaban versus warfarin in patients with atrial fibrillation. *N Engl J Med* 2011;**365**(11):981-92.
47. Negoro S, Masuda N, Takada Y, et al. Randomised phase III trial of irinotecan combined with cisplatin for advanced non-small-cell lung cancer. *Br J Cancer* 2003;**88**(3):335-41.
48. Patel JD, Socinski MA, Garon EB, 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**(34):4349-57.
49. Tunis SL, Faries DE, Nyhuis AW, et al. Cost-effectiveness of olanzapine as first-line treatment for schizophrenia: Results from a randomized, open-label, 1-year trial. *Value in Health* 2006;**9**(2):77-89.
50. 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.
51. Bai X, Liu J, Li L, et al. Adaptive truncated weighting for improving marginal structural model estimation of treatment effects informally censored by subsequent therapy. *Pharm Stat* 2015;**14**(6):448-54.
52. Toh S, Hernandez-Diaz S, Logan R, et al. Estimating absolute risks in the presence of nonadherence: an application to a follow-up study with baseline randomization. *Epidemiology* 2010;**21**(4):528-39.
53. Protocol to CALERIE study available from [https://calerie.duke.edu/files/phase2\\_protocol.pdf](https://calerie.duke.edu/files/phase2_protocol.pdf).
54. Saquib N, Saquib J, Ioannidis JP. Practices and impact of primary outcome adjustment in randomized controlled trials: meta-epidemiologic study. *BMJ* 2013;**347**:f4313.
55. Ioannidis JP, Caplan AL, Dal-Re R. Outcome reporting bias in clinical trials: why monitoring matters. *BMJ* 2017;**356**:j408.
56. Estruch R, Ros E, Salas-Salvadó J, et al. Retraction and Republication: Primary Prevention of Cardiovascular Disease with a Mediterranean Diet. *N Engl J Med* 2013;**368**:1279-90. *New England Journal of Medicine* 2018;**378**(25):2441-42.
57. Ioannidis JPA, Caplan AL, Dal-Ré R. Outcome reporting bias in clinical trials: why monitoring matters. *BMJ* 2017;**356**.
58. Given B, Given CW, Sikorskii A, et al. Analyzing symptom management trials: the value of both intention-to-treat and per-protocol approaches. *Oncol Nurs Forum* 2009;**36**(6):E293-302.
59. Porta N, Bonet C, Cobo E. Discordance between reported intention-to-treat and per protocol analyses. *J Clin Epidemiol* 2007;**60**(7):663-9.
60. Schiffner R, Schiffner-Rohe J, Gerstenhauer M, et al. 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**(6):1154-60.
61. 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* 2015;**350**:h2445.

62. Alshurafa M, Briel M, Akl EA, et al. Inconsistent definitions for intention-to-treat in relation to missing outcome data: systematic review of the methods literature. *PLoS One* 2012;**7**(11):e49163.
63. Bassler D, Briel M, Montori VM, et al. Stopping randomized trials early for benefit and estimation of treatment effects: systematic review and meta-regression analysis. *JAMA* 2010;**303**(12):1180-7.
64. Bell ML, Fiero M, Horton NJ, et al. Handling missing data in RCTs; a review of the top medical journals. *BMC Med Res Methodol* 2014;**14**:118.
65. Del Re AC, Maisel NC, Blodgett JC, et al. Intention-to-treat analyses and missing data approaches in pharmacotherapy trials for alcohol use disorders. *BMJ Open* 2013;**3**(11):e003464.
66. Hernan MA, Robins JM. Per-Protocol Analyses of Pragmatic Trials. *N Engl J Med* 2017;**377**(14):1391-98.
67. Lopes RD, Alexander JH, Al-Khatib SM, et al. Apixaban for reduction in stroke and other Thromboembolic events in atrial fibrillation (ARISTOTLE) trial: design and rationale. *Am Heart J* 2010;**159**(3):331-9.
68. Hennekens CH, Eberlein K. A randomized trial of aspirin and beta-carotene among U.S. physicians. *Prev Med* 1985;**14**(2):165-8.
69. Patel JD, Bonomi P, Socinski MA, 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.
70. 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**(1):61-109.
71. Buring JE, Hennekens CH. The Women's Health Study: Summary of the study design. *J Myocardial Ischemia* 1992;**4**:27-29.

## Figures

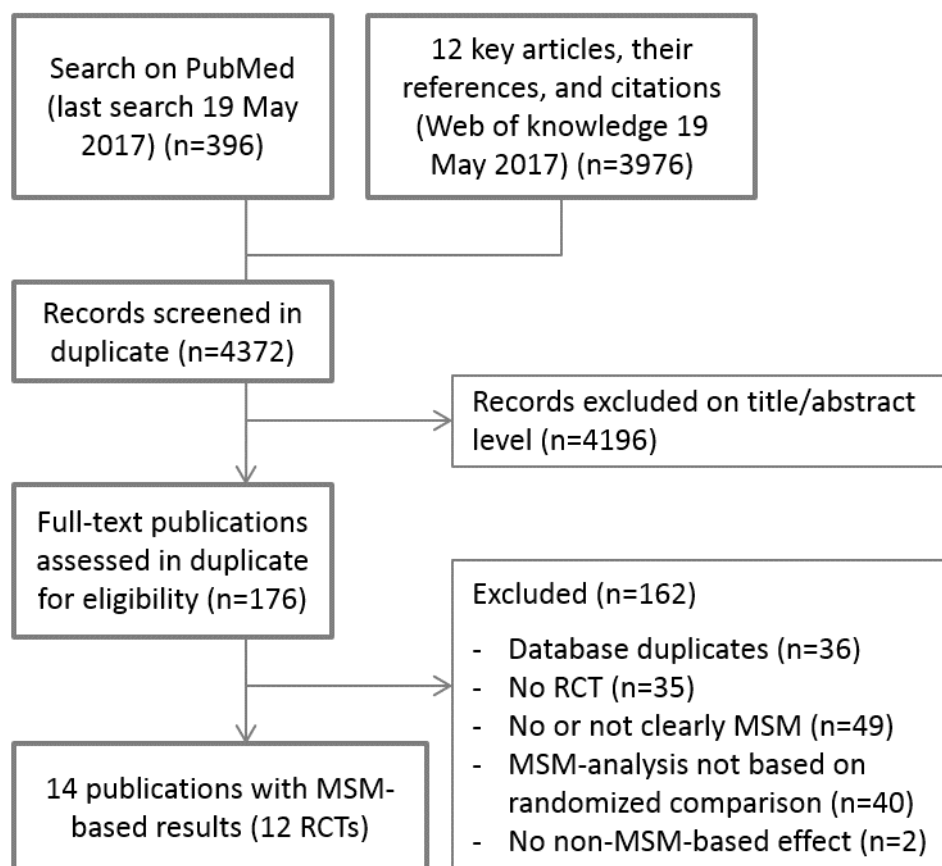


Figure 1 Study flow

MSM: marginal structural models; RCT: randomized controlled trial

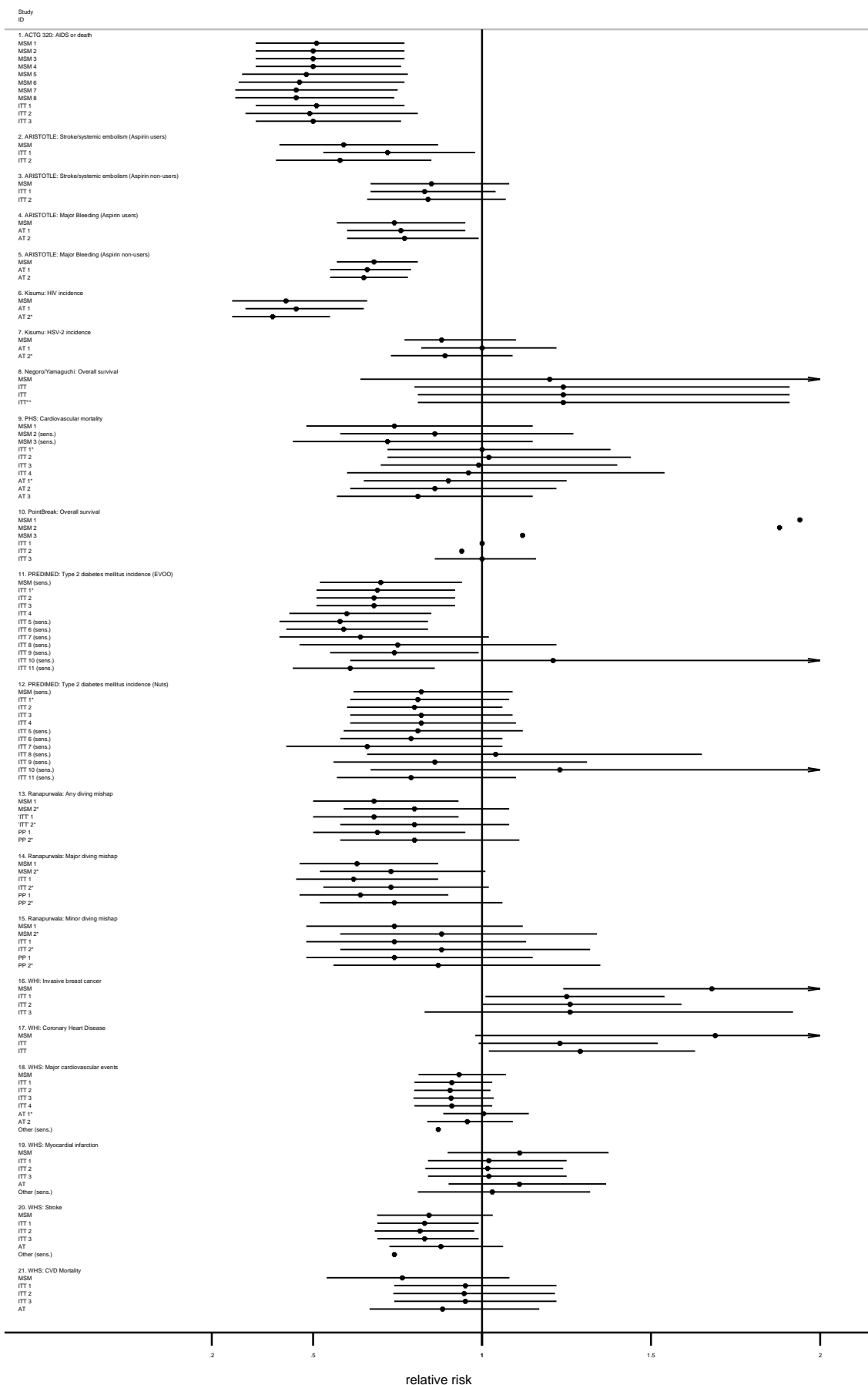


Figure 2 Overview of results from different analyses for the same clinical question (population, intervention, control, outcome) for all 24 clinical questions reported in the main publication and the publication with MSM-results. Circles indicate effect estimates, lines 95% confidence intervals (in some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

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

## Tables

**Table 1 Characteristics of included studies**

<b>RCT</b>	<b>No. Randomized</b>	<b>Patients' Condition</b>	<b>Intervention and control</b>	<b>Outcomes with MSM-based results Analytic approach (n)</b>	<b>Total number of pertinent clinical questions</b>
ACTG 320* <sup>29 43</sup>	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 <sup>28 46 67</sup>	18201 (using aspirin at BL: 5632)	Atrial fibrillation (in aspirin users and non-users)	Apixaban vs Warfarin	Stroke or systemic embolism (primary, subgroups only) MSM (1) ITT (2) Major bleeding MSM (1) As treated (2)	6
	18201 (Not using aspirin at BL: 12569)			Stroke or systemic embolism (primary, subgroups only) MSM (1) ITT (2) Major bleeding MSM (1) As treated (2)	6
CALERIE <sup>34 35</sup>	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 * <sup>25 26 42</sup>	2784	Uncircumcised, HIV-negative young men	Immediate vs delayed circumcision	HIV incidence (primary) MSM (1) As treated (2)	6



				Herpes simplex virus 2 incidence MSM (1) As treated (2)	
Negoro / Yamaguchi *** <sup>36 47</sup>	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 * <sup>30 32 68</sup>	22071	Male physicians	Aspirin vs placebo	Cardiovascular mortality (primary) MSM (3) ITT (4) As treated (3)	10
PointBreak <sup>48 51</sup> <sub>69</sub>	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
PREDIMED * <sup>27</sup> <sub>40</sub>	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 <sup>45</sup>	1660 (70 randomized units)	Recreational scuba divers	Checklist vs no checklist	Any diving mishap (primary) MSM (2) ITT (2) Per protocol (2)	18

				Major diving mishaps MSM (2) ITT (2) Per protocol (2) Minor diving mishaps MSM (2) ITT (2) Per protocol (2)	
Tunis / Faries 37 49	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 * 33 44 70	16608	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 * 31 50 71	39876	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	25

---

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 (effect estimates 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 study compared of 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<sup>-2</sup>), we do not present the third arm (100 mg m<sup>-2</sup>).

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**

<b>Clinical question</b>	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 effect estimates with MSM)*#	Ratio of the relative risks with MSM and ITT* (x-fold difference between effect estimates)	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
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 estimate)	NA (no ITT effect estimate)	NA (no ITT effect estimate)	NA (no ITT effect estimate)
ARISTOTLE (Aspirin users): Major Bleeding	1.04	NA	NA	NA (no ITT effect estimate)	NA (no ITT effect estimate)	NA (no ITT effect estimate)	NA (no ITT effect estimate)
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 estimate)	NA (no ITT effect estimate)	NA (no ITT effect estimate)	NA (no ITT effect estimate)
Kisumu: HSV-2 incidence	1.14	NA	NA	NA (no ITT effect estimate)	NA (no ITT effect estimate)	NA (no ITT effect estimate)	NA (no ITT effect estimate)
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 effect estimates 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

## 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. SYNTHESIS 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, Borrás 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
(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 on the motivation to use MSM	Statements on pre-specification of MSM
ACTG 320 <sup>29</sup>	Time-dependent confounding Non-compliance	"Inverse probability-of-censoring weights for the correction of <b>time-varying noncompliance</b> in the effect of randomized highly active antiretroviral therapy on incident AIDS or death"	Not identified
ARISTOTLE <sup>28</sup>	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 <b>adjust for potential time-dependent confounders</b> 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 <b>accounted for whether a patient was actually taking aspirin</b> at the time of their bleeding event resulted in similar findings"	Not identified

CALERIE <sup>35</sup>	Time-dependent confounding Non-adherence	<p>“Application of the marginal structural model to account for suboptimal <b>adherence</b> in a randomized controlled trial”</p> <p>“First, stepwise linear regression was used to model the observed percent weight loss, while stepwise logistic regression model was applied to model early <b>discontinuation</b> from the intervention.”</p> <p>“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 <b>time-dependent covariates</b> must be considered.”</p> <p>“However, <b>adherence</b> is an endogenous variable and satisfies the definition of a <b>time-dependent confounder</b>.”</p> <p>“The dataset is then analyzed using a logistic regression model including a term for the time interval as well as the fixed and <b>time-dependent covariates</b> of interest.”</p> <p>“Thus, the causal model will be used to address the following types of questions:</p> <ul style="list-style-type: none"> <li>• What is the estimated treatment difference adjusting for</li> </ul>	<p>“To address these <b>mechanistic questions</b>, causal analysis [257-259] will also be applied – specifically, the Marginal Structural Mean (MSM) of Robins and colleagues [260,261].” (from protocol, <a href="https://calerie.duke.edu/files/phase2_protocol.pdf">https://calerie.duke.edu/files/phase2_protocol.pdf</a>)</p>
-----------------------	---	--	--



		<p>imperfect participant <b>adherence</b> to the protocol?</p> <ul style="list-style-type: none"> <li>• What is the estimated treatment difference adjusting for participant <b>drop-out</b> during the study?</li> <li>• What is the estimated treatment difference adjusting for <b>protocol violations</b> during the study?" (from protocol)</li> </ul>	
Kisumu <sup>25 26</sup>	Time-dependent confounding and loss to follow-up	<p>"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."</p> <p>"For our marginal structural approach, we generated the above described stabilized IPTW and stabilized inverse-probability-of-censoring-weights (IPCWs) <b>to account for time-dependent confounding and loss to follow-up.</b>"</p>	Not identified
Negoro/Yamaguchi <sup>36</sup>	Time-dependent confounding Secondline treatment	<p>"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 <b>differential proportions of second-line treatment.</b> In this paper, we deal with clinical</p>	Not identified

		<p>application of the two models in detail.”</p> <p>“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 <b>time-dependent confounders</b> which satisfy the condition (i) and (ii) introduced in Section 3.1.”</p>	
PHS <sup>30</sup>	Time-dependent confounding	<p>"The authors used a marginal structural model with <b>time-dependent inverse probability weights</b> to estimate the underlying causal effect of aspirin on cardiovascular mortality."</p> <p>"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 <b>adjustment for time-varying covariates</b> in time-dependent models."</p>	Not identified
PointBreak <sup>51</sup>	Time-dependent confounding	<p>"Marginal structural models (MSMs) have been applied to estimate causal treatment effects even in the presence of <b>time-dependent confounders</b>."</p>	Not identified

PREDIMED <sup>27</sup>	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."	Not identified
Ranapurwala <sup>45</sup>	Non-adherence	"Marginal structural models were used <b>to account for non-adherence.</b> "	Not clear: "Having multiple outcomes and multiple dives in the data, we will conduct survival analysis using Poisson regression model or generalized estimating equation (GEE) models." (from protocol kindly provided by authors)
Tunis/Faries <sup>37</sup>	Treatment switching Time-dependent confounding	<p>"Various methods of <b>eliminating the switching</b>, such as epoch analyses and on-drug subset analyses, along with use of marginal structural models generated reasonably consistent non-zero treatment effect estimates."</p> <p>"The MSM retains the repeated measures structure of the data and directly <b>addresses time-varying covariates</b> 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)."</p>	Not identified

		<p>"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 <b>treatment switching</b> and <b>time-varying confounders</b>."</p>	
WHI <sup>33 52</sup>	Time-dependent confounding Non-adherence	<p>"Inverse probability weighting of marginal structural models has been used <b>to adjust for nonadherence</b>, but most studies have provided only relative measures of risk."<sup>52</sup></p> <p>"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 <b>time-varying selection bias</b> due to artificial censoring."<sup>52</sup></p> <p>"<b>Adherence-adjusted</b> hazard ratios and CHD-free survival curves estimated through inverse probability weighting."<sup>33</sup></p>	Not identified
WHS <sup>50</sup>	Time-dependent confounding Non-compliance	<p>"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</p>	Not identified

---

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

Included RCT	Method of estimation	Model	Mean range and handling of weights
ACTG 320 <sup>29</sup>	“Using data from the AIDS Clinical Trial Group (ACTG) 320 study reported by Hammer et al. [ 2 ], we apply RF <b>inverse probability-of-censoring weights</b> [ 10 ] to estimate the average causal effect of highly active ART versus combination ART on time to AIDS or death correcting for noncompliance and extend the method to account for non-proportional hazards.”	“Both the conditional probabilities for numerator and denominator were fit using <b>pooled logistic regression models</b> [ 21 ] for the discrete-time hazard of censoring, specifically, [...]”	" <b>Truncating</b> the inverse probability-of-censoring weights at the 1st and 99th percentiles [ 26 ] did not appreciably change the results"
ARISTOTLE <sup>28</sup>	“[...] weighting was used to adjust for time-varying confounders [...]”	“This approach involves a <b>Cox model</b> where aspirin status is a time-dependent covariate [...]”	N.r.
CALERIE <sup>35</sup>	“The marginal structural model with <b>inverse-probability weighting</b> was implemented using a weighted generalized estimating equation model.”	“The marginal structural model with inverse-probability weighting was implemented using a weighted <b>generalized estimating equation model</b> . Two ancillary models were developed to derive the weights. First, stepwise <b>linear regression</b> was used to model the observed percent weight loss, while stepwise <b>logistic regression</b> model was applied to model early discontinuation from the intervention. From these, participant- and time-specific weights were calculated.”	“The <b>stabilized weight</b> for each participant at each visit, therefore, was the ratio of the cumulative probabilities from excluding and including the time-dependent covariates.”  “The final weight for each CR participant at each visit was the product of the <b>stabilized weights</b> from the two sets of ancillary analyses.”
Kisumu <sup>25 26</sup>	“Cox proportional hazards regression incorporating stabilized <b>inverse probability of treatment and censoring weights</b>	“ <b>Cox proportional hazards regression</b> incorporating stabilized inverse probability of treatment and censoring weights generated	Results of logistic regression models to generate weights are presented for each

	<p>generated through marginal structural modeling was used to estimate the efficacy of MMC on HSV-2 risk.”</p> <p>“Cox proportional hazards regression incorporating stabilized <b>inverse probability of treatment and censoring weights</b> generated through marginal structural modeling, was used to account for potential time-varying confounding and censoring to estimate the efficacy of MMC on HIV risk.”</p>	<p>through marginal structural modeling was used to estimate the efficacy of MMC on HSV-2 risk.”</p> <p>“<b>Cox proportional hazards regression</b> incorporating stabilized inverse probability of treatment and censoring weights generated through marginal structural modeling, was used to account for potential time-varying confounding and censoring to estimate the efficacy of MMC on HIV risk.”</p> <p>“Weights were obtained through fitting a weighted <b>pooled logistic regression model</b> for circumcision status (or censoring) using each study visit as an observation.”</p> <p>“To obtain stabilized IPTWs, we fitted a <b>weighted pooled logistic regression model</b> for circumcision using each 6-month study visit as an observation. The weighted logistic model was fitted using generalized estimating equations assuming an exchangeable correlation structure, with robust variance estimation, to account for correlated observations.”</p>	<p>variable in tables 2 in the respective publication<sup>25 26</sup></p> <p>“Cox proportional hazards regression incorporating <b>stabilized</b> inverse probability of treatment and censoring weights generated through marginal structural modeling was used to estimate the efficacy of MMC on HSV-2 risk.”</p> <p>“Cox proportional hazards regression incorporating <b>stabilized</b> inverse probability of treatment and censoring weights generated through marginal structural modeling, was used to account for potential time-varying confounding and censoring to estimate the efficacy of MMC on HIV risk.”</p>
Negoro/ Yamaguchi <sup>36</sup>	<p>“The parameters of marginal structural models can be consistently estimated using the <b>inverse-probability-of-treatment weighted</b> (IPTW) estimators [10–12].”</p> <p>“Figure 6 shows box plots for the distribution of the estimated probability of each patient having his = her history of second-line radiotherapy (that is, the <b>inverse</b> of <math>\hat{W}_i(t)</math>) and the probability of each patient</p>	<p>“We consider the <b>marginal structural Cox proportional hazards model</b> [...]”</p>	<p>“If we use stabilized weights, <math>\hat{S}_i(t)</math> is <b>ranged</b> from <b>0.069 to 3.253</b>, which is smaller variation compared to non-stabilized weights.”</p> <p>“The distribution of non-stabilized weights is very skewed and there exists some extreme values. On the other hand, the distribution of <b>stabilized</b> weights is symmetric and much less variable than</p>

	uncensored (that is, the <b>inverse</b> of $\hat{W}_i(t)$ ) at an interval of about 3 months.”		non-stabilized one at each time of follow-up. Therefore, we show the result using the stabilized weights.”
PHS <sup>30</sup>	“The authors used a marginal structural model with time-dependent <b>inverse probability weights</b> to estimate the underlying causal effect of aspirin on cardiovascular mortality.”	“ <b>Pooled logistic regression analyses</b> predicting aspirin use and censoring in each year were used to estimate the weights.”	<p>“We also fit models for the probability of remaining uncensored by either death due to other causes or by the end of study to obtain <b>estimates of the censoring weights <math>CW_i(t)</math>, which tended to be near 1.</b>”</p> <p>See also weight in Appendix figure 1 in the publication<sup>30</sup></p> <p>“APPENDIX FIGURE 1. Diagram of a hypothetical population of 200,000 persons randomized to aspirin or placebo use at baseline. <math>ASA_1</math> and <math>ASA_2</math> represent aspirin use at times 1 and 2, respectively, and an overbar represents an absence of aspirin use or myocardial infarction (MI). N, frequency distribution; <b>SW, stabilized weights.</b>”</p>
PointBreak <sup>51</sup>	“We refer to MSMs using the proposed approach to estimate weights as <b>adaptive truncating longitudinal inverse-probability weighting</b> in marginal structural models (AtMSM).”	<p>“In the current article, we proposed a new method to estimate MSM weights by adaptively truncating the longitudinal inverse-probability computations, referred as adaptive truncated marginal structural model (AtMSM).”</p> <p>“We focus on <b>marginal structural Cox proportional hazards models</b> in this article, although our method can be applied to other MSMs.”</p>	<p>“For the majority of patients, the log range of their weight function is between 3 and 3.”</p> <p>“Truncating the weight at the last time point may be too late to keep weights in a reasonable range, so we calculated the weights and <b>truncated</b> them iteratively at each time point (e.g., day).”</p> <p>“The <b>truncation approach</b> can be viewed as a special case of kernel smoothing [14], which is a commonly used as non-</p>



			<p>parametric estimation approach to improve efficiency.”</p> <p>“This led to the <b>extremely large weights</b> for these patients. By <b>applying the AtMSM</b>, the resulting <b>estimators are more realistic and stable</b>, with a smaller variance, compared to standard and augmented MSMs.”</p>
PREDIMED <sup>27</sup>	<p>“We applied marginal structural models with <b>inverse probability weighting</b> (21, 22), using the variables with small between-group imbalances to compute the weights (we computed the stabilized weights using logit, the postestimation command predict, and then generate).”</p>	<p>“<b>Cox regression models</b> (stset, stcox) were fitted to assess hazard ratios (HRs) for diabetes for the 2 Mediterranean diet groups in comparison with the control group.”</p>	<p>“We applied marginal structural models with inverse probability weighting (21, 22), using the variables with small between-group imbalances to compute the weights (we computed the <b>stabilized</b> weights using logit, the postestimation command predict, and then generate).”</p>
Ranapurwala <sup>45</sup>	<p>“Marginal structural models were used to account for non-adherence. <b>Inverse probability weights</b> for adherence were calculated for all intervention group participants, whereas the control group participants were given a weight of one.”</p>	<p>“The adherence probabilities were derived using <b>logistic regression</b>. Crude and adjusted <b>Poisson regression models</b> were then weighted using these inverse probability weights.”</p>	<p>From figure 1 in the publication<sup>45</sup>:</p> <p>Intervention group: <b>weights=617, weight range 0.85 to 1.49</b></p> <p>Control group: <b>weights=426, weight range 1 to 1</b></p>
Tunis/Faries <sup>37</sup>	<p>“The MSM analysis required 3 separate models: one for treatment selection probability (either continuation or switching), one for missing data probability, and a third for repeated measures analysis using <b>inverse probability re-weighting</b> from the first two models.”</p>	<p>“As the first step in the MSM analysis, <b>binomial and multinomial logistic models</b> were utilized to compute weights for treatment selection and censoring for each patient visit in the study.”</p> <p>“Treatment selection weights for the MSM analysis were estimated using a <b>multinomial model</b>—with treatment as the dependent variable.”</p> <p>“Causal treatment effects were then estimated using a weighted <b>repeated</b></p>	<p>Table 5, “Summary of <b>MSM weight</b> (censoring and treatment selection) models”, in the publication<sup>37</sup> reports degrees of freedom, Chi-square and p-values for the censoring and the treatment selection model</p> <p>“<b>Stabilized</b> weights for missing data were estimated using a logistic regression model, with a binary flag for remaining-in-study as the outcome variable.”</p>

	<b>measures model with generalized estimating equations</b> (PROC GENMOD, SAS) and an exchangeable correlation matrix."		
WHI <sup>33 52</sup>	<p>"In this study, we used <b>inverse probability weighting</b> to estimate both absolute and relative measures of risk of invasive breast cancer under full adherence to the assigned treatment in the Women's Health Initiative estrogen-plus-progestin trial." <sup>51</sup></p> <p>"We estimated the <b>inverse probability weights</b> by fitting, separately for each group, a logistic regression model to estimate each participant's probability of receiving hormone therapy during each follow-up year and a linear regression model to estimate each participant's density (assumed to be normal) of receiving their actual proportion of pills taken (log-transformed) among those with nonzero use during that year (9, 26)." <sup>32</sup></p>	<p>"We estimated these quantities by fitting, separately for each arm, (1) a <b>logistic regression model</b> to estimate the probability of receiving any hormone therapy, and (2) a <b>linear regression model</b> that assumed independent normal errors with constant variance to estimate the density of receiving the log-transformed proportion of pills taken among those with non-zero use during that year." <sup>51</sup></p> <p>"We estimated the inverse probability weights by fitting, separately for each group, a <b>logistic regression model</b> to estimate each participant's probability of receiving hormone therapy during each follow-up year and a <b>linear regression model</b> to estimate each participant's density (assumed to be normal) of receiving their actual proportion of pills taken (log-transformed) among those with nonzero use during that year (9, 26)." <sup>32</sup></p>	<p>"The <b>mean</b> of the estimated stabilized inverse probability <b>weights</b> for adherence adjustment was <b>1.00 (standard deviation = 0.29)</b>" <sup>51</sup></p> <p>"The <b>mean</b> of the estimated stabilized inverse probability weights for adherence adjustment was <b>1.00 (SD, 0.30)</b>." <sup>32</sup></p> <p>"To improve statistical efficiency, the weights were <b>stabilized</b> [...]." <sup>51</sup></p> <p>"Results were similar under all these models although, when estimating conditional variances, we had to <b>restrict the analysis to a subset of the covariates to avoid extreme stabilized weights</b>." <sup>51</sup></p> <p>"The mean of the estimated <b>stabilized</b> inverse probability weights for adherence adjustment was 1.00 (SD, 0.30)." <sup>32</sup></p>
WHS <sup>50</sup>	<p>"These two sets of weights were multiplied and accumulated over time to form the <b>inverse probability weights</b>."</p>	<p>"Treatment weights for the major CVD endpoint were constructed using the <b>logistic regression results</b> shown in Table 1, which served as the denominators of the weights. The numerators were created from <b>similar models</b> including year, age, race and previous aspirin use only."</p>	<p>"After multiplication of treatment and censoring weights, the <b>mean</b> weight was <b>1.005</b>, with a <b>median</b> of <b>0.999 (interquartile range of 0.967–1.009)</b>. The weights were truncated at the 0.01th and 99.99th percentiles, representing <b>values of 0.022 and 6.914</b>."</p> <p>"To <b>stabilize</b> the weights and reduce their variability, the numerator of the weights consisted of predicted probability from a</p>

---

second logistic model for observed aspirin as a function of past aspirin use and a subset of baseline variables, without including intervening factors [17]."

"The weights were **truncated** at the 0.01th and 99.99th percentiles, representing values of 0.022 and 6.914."

---

IPW= inverse probability weighting; N.r.=not reported

## Webappendix 4

### Details of study results from main MSM-based and ITT-based analyses for main clinical questions 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 (Tables 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 <sup>28</sup>. 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 <sup>46</sup>.

---

#### **Physicians' Health Study (PHS)**

The Physicians' Health Study (PHS) explored the effects of aspirin compared to placebo on cardiovascular mortality (Tables Table 1). The main MSM-based result was a relative risk (RR) of 0.74 (95% CI 0.48 to 1.15) favoring aspirin <sup>30</sup>. The main ITT analysis was RR = 0.96 (95% CI 0.6 to 1.54) <sup>32</sup>.

---

#### **Women's Health Study (WHS)**

The Women's Health Study (WHS) studied the effects of aspirin vs placebo on major cardiovascular events (Tables Table 1). The main MSM-based result was a RR of 0.93 (95% CI 0.81 to 1.07) <sup>50</sup>. The main ITT effect estimate was a RR of 0.91 (95% CI 0.8 to 1.03) <sup>31</sup>.

---

#### **Ranapurwala**

The study by Ranapurwala et al. assessed the effect of a pre-dive checklist on diving mishaps (Tables

---

---

Table 1). The main MSM-based result was a RR of 0.68 (95% CI 0.5 to 0.93)<sup>45</sup>. The main ITT analysis was a RR of 0.8 (95% CI 0.58 to 1.08)<sup>45</sup>.

---

### **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 (Tables

Table 1). MSM-based effect estimates 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<sup>27</sup>. The according main ITT effect estimates were HR 0.69 (95% CI 0.51 to 0.92) and 0.81 (95% CI 0.61 to 1.08), respectively<sup>27</sup>.

---

### **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 (Tables

Table 1). The main MSM-based effect estimate was a HR of 1.69 (95% CI 0.98 to 2.89)<sup>33</sup> and the main ITT effect estimate was a HR of 1.29 (95% CI 1.02 to 1.63)

<sup>44</sup>.

---

### **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 (Tables

Table 1). The main MSM-based result was a HR of 0.45 (95% CI 0.27 to 0.74)<sup>29</sup>. The main ITT result was an HR of 0.5 (95% CI 0.33 to 0.76)<sup>43</sup>.

---

### **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 (Tables

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)<sup>51</sup>. The main ITT result was a HR of 1.00 (95% CI 0.86 to 1.16)<sup>48</sup>.

---

#### **Tunis / Faries**

The RCT reported by Tunis et al.<sup>49</sup> and Faries et al.<sup>37</sup> 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 (Tables

Table 1). The main MSM-based estimated treatment difference (olanzapine - conventional) was 1.9 (95% CI 0.5 to 3.3)<sup>37</sup> and the corresponding ITT estimated treatment difference was 0.2 (95% CI -1.8 to 2.1)<sup>49</sup>.

---

#### **Negoro / Yamaguchi**

The RCT reported by Negoro et al.<sup>47</sup> and Yamaguchi et al.<sup>36</sup> explored the effect of irinotecan hydrochloride versus cisplatin in patients with lung cancer on overall survival (Tables

Table 1). An MSM-based effect estimate 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)<sup>36</sup>. The main ITT analysis was HR = 1.24 (95% CI 0.81 to 1.91)<sup>47</sup>.

---

#### **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 (Tables

---

---

Table 1). We used resting metabolic rate for the main clinical question. The main MSM-based mean difference was -36 (95% CI not reported) <sup>35</sup> and the corresponding ITT mean difference was -64 (95% CI not reported) <sup>34</sup>.

---

### **Kisumu RCT**

The Kisumu RCT assessed the effect of immediate versus delayed circumcision on HIV incidence (Tables

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) <sup>26</sup>. Hence, this trial was not used in the main comparison of MSM vs ITT.

---