Addressing Systematic Missing Data in the Context of Causally Interpretable Meta-analysis

David H. Barker[1,2] · Ruofan Bie[3] · Jon A. Steingrimsson[3]

Abstract

Evidence synthesis involves drawing conclusions from trial samples that may differ from the target population of interest, and there is often heterogeneity among trials in sample characteristics, treatment implementation, study design, and assessment of covariates. Stitching together this patchwork of evidence requires subject-matter knowledge, a clearly defined target population, and guidance on how to weigh evidence from different trials. Transportability analysis has provided formal identifiability conditions required to make unbiased causal inference in the target population. In this manuscript, we review these conditions along with an additional assumption required to address systematic missing data. The identifiability conditions highlight the importance of accounting for differences in treatment effect modifiers between the populations underlying the trials and the target population. We perform simulations to evaluate the bias of conventional random effect models and multiply imputed estimates using the pooled trials sample and describe causal estimators that explicitly address trial-to-target differences in key covariates in the context of systematic missing data. Results indicate that the causal transportability estimators are unbiased when treatment effect modifiers are accounted for in the analyses. Results also highlight the importance of carefully evaluating identifiability conditions for each trial to reduce bias due to differences in participant characteristics between trials and the target population. Bias can be limited by adjusting for covariates that are strongly correlated with missing treatment effect modifiers, including data from trials that do not differ from the target on treatment modifiers, and removing trials that do differ from the target and did not assess a modifier.

Keywords Systematic missing data · Causal inference · Individual participant data

Using data from clinical trials to understand what interventions work best for which populations requires combining evidence from trials that sampled from different populations or subpopulations, evaluated different treatment implementations, used different study designs, and used different assessments of outcomes and covariates. Between-trial differences contribute to observed heterogeneity in treatment effects that can obscure decisions about which intervention may work best for a given target population (Brincks et al., 2018). Application of causal methods to evidence synthesis has provided guidance on the conditions required to generalize causal inference from trial samples to a target population (Dahabreh et al., 2019a, b, c; Markozannes et al., 2021; Stuart et al., 2001; Susukida et al., 2017), thus helping clarify the sources of treatment heterogeneity by addressing differences among trial and target populations. If all trials sample from the same target population, then conventional meta-analytic approaches are sufficient for generalizing from trials to the target. If, however, the trials sample from different populations or if those populations differ from the target, additional methods are required (Markozannes et al., 2021). Transportability methods refer to approaches used to transport findings from trial populations to target populations with different distributions of participant characteristics. The methods define identifiability conditions (i.e., assumptions needed for doing causal inference with the observed data); use subject-matter knowledge to evaluate the feasibility of the conditions; and

✉ David H. Barker
dbarker@lifespan.org

1 Department of Psychiatry and Human Behavior, The Warren Alpert Medical School of Brown University, Providence, RI, USA

2 Bradley Hasbro Children's Research Center, Providence, RI, USA

3 Department of Biostatistics, Brown University, Providence, RI, USA

use covariate and outcome data from trials, covariate data from a sample of the target population, and statistical models to produce unbiased estimates of treatment effects in the target population (Barker et al., 2021; Dahabreh et al., 2020a, b; Steingrimsson & Yang, 2019).

Systematic missing data is a challenge for methods utilizing individual participant data. Systematic missing data occurs when all trials do not measure the same set of variables. Missing a key covariate can, but does not always, introduce bias into the estimation of the treatment effect in the target population. Understanding the conditions that lead to bias can help reduce bias by informing decisions about trial inclusion and informing analytic decisions around covariate inclusion.

There has been work addressing systematic missing data in meta-analysis of individual participant data (Burgess et al., 2013; Jolani et al., 2015; Kunkel & Kaizar, 2017; Resche-Rigon et al., 2013; Resche-Rigon & White, 2018; Siddique et al., 2015, 2018). Results indicate that multiple imputation that accounts for the multilevel structure of participants nested with trials results in minimal bias when estimating the treatment effect in the pooled trials sample. Importantly, this previous work defined the target population using the pooled trials sample, which is only well defined when trials are randomly sampled from the same population (Dahabreh et al., 2019a, b, c). When the populations underlying the trials differ, the pooled sample is a mixture of populations weighted by trial sample size resulting in a sample from a hard to characterize population that does not have a clear causal interpretation (Markozannes et al., 2021).

Recently, the identifiability conditions for systematic missing data in the context of transportability analysis from multiple trials to a target have been articulated (Steingrimsson et al., 2023). In this manuscript, we will review the identifiability conditions necessary for causal interpretation of meta-analytic results, describe the additional condition needed for unbiased estimation in the presence of systematic missing data, and describe causal estimators that can be used to address systematic missing data. We used Monte Carlo simulations to compare these causal estimators with conventional random effect and imputation models while varying the strength of treatment effect modifiers, size of the trial-to-target differences, and pattern of systematic missing data.

Identifiability Conditions

To transport a causal parameter (e.g., average treatment effect, potential outcome mean) from trial samples to a target population, certain identifiability conditions must be met (Barker et al., 2021; Dahabreh et al., 2020a, b; Steingrimsson & Yang, 2019). We assume that the analyst has access to individual participant data on treatment assignment A, the same outcome

Y, and baseline covariates X from a set of trials $S \in \{1, \ldots, m\}$. We also assume that we have access to a random sample of baseline covariates from the target population, where an observation coming from the target population is denoted by $S = 0$. Let Y_i^a be the potential outcome if individual i were assigned to treatment a. The potential outcome mean in the target population is thus denoted as $\psi(a) = E[Y^a | S = 0]$ and the average treatment effect contrasting treatments a and a' as $\psi(a) - \psi(a\prime)$.

The identifiability conditions required for causal inference in the target population are as follows:

(1) Consistency: If $A_i = a$, then $Y_i^a = Y_i$ for every observation in the collection of trials and the target population. This implies that all participants who received treatment a received the same version of the treatment or that variation in treatment is not informative (Dahabreh et al., 2019a, b, c; Hernán & VanderWeele, 2011; VanderWeele, 2009), there is no interference, and there is no effect of trial engagement on the outcome.

(2) Within trial treatment exchangeability: For all treatments a and all trials s, $Y^a \perp\!\!\!\perp A | (X, S = s, S \neq 0)$. This assumption states that potential outcomes are independent of treatment assignment conditioned on baseline covariates and holds when treatment assignment is randomized or when there is no unmeasured confounding for non-randomized treatments.

(3) Positivity of treatment assignment: For every treatment a, trial s, and covariate pattern x, $Pr[A = a | X = x, S = s] > 0$. This assumption states that all trial participants have a nonzero probability of being assigned to each treatment in trial s. This condition holds in randomized trials but may be violated in non-randomized designs.

(4) Trial to target exchangeability: For all treatments a, $Y^a \perp\!\!\!\perp S | X$. This assumption states that the potential outcomes are independent from trial or target assignment conditioned on baseline covariates.

(5) Positivity of trial participation: $Pr[S \neq 0 | X = x] > 0$ for every pattern of covariates x that can occur in the target population. This assumption states that every covariate pattern in the target population can occur in at least one of the trial samples.

The first three assumptions are standard assumptions made when analyzing data from randomized studies, and the last two are additional assumptions required for transportation from trials to a target population. Informally, the baseline covariates that need to be adjusted for when considering the average treatment effect are those that modify treatment response and differ between trials and the target population.

The additional *identifiability condition required for systematic missing data* is that for every set of trials that shares the same pattern of systematic missing data, the missing

covariates are independent of the outcome given observed covariates. This is akin to a missing at random assumption for systematic missing data. The mathematical justification for this assumption has been presented elsewhere (Steingrimsson et al., 2023). Although the additional assumption is for all trials that share the same pattern of systematic missing data, in practice, it is likely that a trial-by-trial evaluation of the identifiability conditions will be more feasible. This work applies to situations where all covariates are collected in the target population, that is, systematic missing data is only in the trial samples.

The implication of this assumption is that each trial can be evaluated separately and use different sets of covariates to satisfy identifiability conditions. For example, suppose that we have two trials, and that treatment response is dependent on symptom severity. Furthermore, suppose that symptom severity was assessed in the first but not the second trial. Unbiased estimates from the first trial could be generated regardless of whether the distribution of symptom severity differs between the trial and target population because severity was assessed in the trial and target and can be adjusted in the estimation process. For the second trial, unbiased estimation is only possible when the trial and target have the same distribution of symptom severity or when the observed covariates account for the link between symptom severity and treatment response. When there are differences in symptom severity between the trial and target, including the second trial would introduce bias into the estimate of the treatment effect in the target population. Similarly, consider that both trials differed from the target population in symptom severity, but that symptom severity was assessed using different instruments in each trial. Presuming that both instruments are available in the target, both trials could account for symptom severity during estimation and produce unbiased estimates while using different measures of severity.

Estimation with Systematic Missing Data

Transporting the average treatment effect from trials to the target population requires individual participant data from the trials including treatment effect modifiers, outcomes, and treatment assignment. It also requires a random sample of covariates from the target population. The approach then builds statistical models that account for the differences in the distribution of treatment effect modifiers between trials and the target population. There are three types of causal estimators that have been developed for use with individual participant data (Steingrimsson et al., 2023): (1) a weighting estimator that models the conditional probability of selection to trial versus target, (2) an outcome model-based estimator (g-formula) that builds a prediction model of the outcome for each treatment condition and then uses the model to predict the outcome in the target population, and (3) a double robust estimator that uses both weighting and an outcome models and is unbiased if either of the two models is correctly specified. The nuisance or working models (i.e., predicting trial selection, predicting treatment assignment, predicting outcome) used for these estimators can be fit using conventional generalized linear models or using more flexible data adaptive approaches such as regularized regression or random forests, but for the more flexible approaches, data splitting is needed for conducting inference (Chernozhukov et al., 2018). Estimators are applied to each trial, or set of trials with the same missing data pattern, producing transported estimates of the treatment effect in the target population for each trial. The combined estimate is obtained by averaging over different estimates weighted by trial sample size. Other choices of weights are possible, but simulations show that the choice of weights has limited impact on performance (Steingrimsson et al., 2023). Standard errors and confidence intervals can be calculated using either sandwich variance estimators or nonparametric bootstrap.

Methods

Simulations were generated using the R package simstudy v0.5.1 and analyzed using the stats v 4.2.0, lme4 v1.1–27.1, and mice v3.13.0 packages. Annotated R code for the simulations and analyses is included in supplemental materials. Data was generated for 20 trials each of size 300 ($n = 6000$) with the target sample being of size 500. Six standard normal covariates were generated ($\times 1 – \times 6$) with weaker correlations ($r = 0.3$) among $\times 1, \times 4$, and the others and stronger correlations ($r = 0.6$) among $\times 2, \times 3, \times 5$, and $\times 6$. The covariates $\times 4 – \times 6$ were discretized into binary variables with a 50/50 split. For all simulations, $\times 3$ and $\times 6$ were used as treatment effect modifiers.

Trial and Population Selection

An indicator for being in the set of samples from the trials versus in the target sample was simulated using a logistic regression model with moderators ($\times 3$ and $\times 6$) as predictors. Assignment to individual trials was generated using a twostep process to ensure that half of the trials were systematically distinct from the other half while also ensuring that on average the trials sample differed from the target by the desired amount. The pooled trials sample ($n = 6000$) was separated into two groups using logistic regression with moderators as predictors, and then trial assignment was evenly distributed among both halves of the pooled trials sample. A between trial-level random variable was added for all covariates.

Treatment and Outcome Generation

For each trial, treatment assignment was marginally randomized using a 1:1 ratio. The outcome variable was simulated in the metric of standardized difference score (d) with moderators ($\times 3$ and $\times 6$) interacting with treatment assignment. We also added a random effect for treatment to help simulate sources of treatment heterogeneity other than sample characteristics (e.g., treatment implementation).

Systematic Missing Data

A subset of trials was selected to have systematic missing data and three patterns of missing data were simulated. First, both moderators ($\times 3$ and $\times 6$) were missing. Second, half of the subset had moderator $\times 3$ set to missing, while the other half had $\times 6$ set to missing. The third pattern set non-moderators $\times 1$ and $\times 4$ to missing.

Simulation Scenarios

The simulation setup allowed us to manipulate the size of covariate differences between target and trial samples, the strength of treatment effect modification, and the proportion of trials with systematic missing data. The setup also had several features to help illustrate use in real-world data. First, we included systematic differences between trial samples as well as random variables for covariates; trial samples can thus be conceptualized as being sampled from two populations that systematically differ from the target and from each other. We set the systematic difference between the two trial subsets to be moderate (odds ratio [OR] = 2.07; mean observed OR across simulations 1.99 [standard deviation 0.19]) and the amount of between-trial sample variability on covariates to 10% (observed intra-class correlations 0.13 [0.06]). Second, we included a random variable for the average treatment effect which simulates heterogeneity among treatment effects not related to sample characteristics. This variability was also set to 10% (observed intra-class correlation 0.12 [0.04]). Third, we included both normally distributed and binary covariates. Finally, we included covariates with differing levels of correlation with the treatment modifiers ($r = 0.3$ & 0.6; observed correlations 0.34 [0.03] and 0.61 [0.03] respectively).

To illustrate how the strength of treatment modifiers and trial versus target sample differences work together to affect bias, we varied the strength of moderators ($\times 3$ & $\times 6$) from none to large ($d = 0$ to 0.8; difference between observed and expected moderator strength $\times 3 = 0.00$ [0.04]; $\times 6 = 0.09$ [0.10]) and the size of trial-to-target differences from none to large (OR = 1 to 4.28; difference between observed and expected log odds $\times 3 = 0.03$ [0.08]; $\times 6 = 0.09$ [0.15]). To illustrate performance with systematic missing data, we set the size of the trial-to-target difference to OR = 2.07 and strength of moderators to $d = 0.4$ while varying the proportion of trials with systematic missing data from 0.1 to 0.6 for each of the three missing data patterns.

Estimators Compared

We used multiple estimators to estimate the average treatment effect in the target population. We used the inverse probability weighting [IPW], outcome modeling [OM], and double robust [DR] causal estimators. The nuisance models for the prediction of target vs trial participation and prediction of outcome used generalized linear models with all available covariates. The nuisance model for prediction of treatment within trial was fit without covariates as the treatment was randomized. For the IPW estimator, we used trimmed and normalized weights (Dahabreh et al., 2020a, b). As comparisons, we used two models based on the pooled trial data common in meta-analysis with IPD. The first was a random effect model with random treatment effect, and the second added all covariates centered on the pooled trial data. Both estimators presume that the pooled trial data is representative of the target population and are expected to be biased if the distribution of treatment modifiers differ between the pooled sample and the target population.

For missing data simulations, we used three approaches to addressing the missing data using the causal estimators: (1) fitting the estimators with only weaker estimators (i.e., covariates $\times 1$ and $\times 4$ that were correlated with the missing moderator at $r = 0.3$), (2) fitting the estimators using both stronger and weaker covariates, and (3) using complete case analysis where the trials with systematic missing data were not used in the analysis. As a comparison, we also fit a multiply imputed estimator that used a 2-level marginal imputation model.

Metrics of Comparison

Bias was computed as the difference between the observed treatment effect in the target and the estimated treatment effect generated by each estimator. All treatment estimates were in the metric of d. We also examined the standard errors for the estimators. For the random effect and imputed models, standard errors were averaged across simulations to retain adjustments for nesting and imputation. For the causal estimators, we used the empirical standard error calculated across simulations.

Results

Results from the simulations varying the strength of the treatment effect moderators while holding the trial-to-target differences constant at none, small (OR = 1.44), medium

(OR = 2.48), and large (OR = 4.28) are presented in Fig. 1. These results showed that all estimators were unbiased when there were no differences between target and trial samples or when the strength of the modifier was negligible. As the strength of the modifier increased so did the bias of the random effect models, but only when there is also a trial-to-target difference. Bias ranged from $d = 0.07$ to 0.48 for small, $d = 0.13$ to 1.02 for moderate, and $d = 0.16$ to 1.31 for large differences. In contrast, there was minimal bias for the IPW estimator ($d = 0.01$ to 0.22) and no bias for the OM and DR estimators. The small bias in the IPW estimator was due to the mismatch between the data generating mechanism for the target vs trial selection model and the simulated data as the simulated data included the additional selection model separating the trials sample into two distinct subsamples.

Results from simulations varying the strength of trial-to-target differences while holding constant the strength of moderation at none, small ($d = 0.2$), medium ($d = 0.5$), and large ($d = 0.8$) are presented in Fig. 2. These results mirrored those in Fig. 1 with all estimators showing no bias when there were no trial-to-target differences or when moderator strength was negligible. As the size of the trial-to-target difference increased, bias of the random effect models increased

from $d = 0.07$ to 0.33 for small, $d = 0.16$ to 0.82 for moderate, and $d = 0.25$ to 1.31 for large moderator strength. There was minimal bias for IPW estimator ($d = 0.01$ to 0.20) and no bias for the OM and DR estimators.

Regarding model uncertainty, standard errors increased for all estimators with the strength of the mediators. For the causal estimators, error also increased with the size of the trial-to-target difference, with the increase being larger for the IPW and DR estimators. There were minimal differences between the two random effect models.

Systematic Missing Data

Simulations with systematic missing data used a small to moderate trial-to-target difference (OR = 2.07) and small to moderate strength ($d = 0.4$) of treatment effect moderation. Results for the three versions of the DR estimator (i.e., only weaker covariates, stronger and weaker covariates, and using only trials with complete data) and for the multiply imputed random effect estimator are presented in Fig. 3. We opted to only show DR estimator to simplify the figure. Results indicate that the multiply imputed estimator showed consistent bias of about $d = 0.42$ with increasing

Fig. 1 Bias with increasing moderator strength. Figure depicts bias and standard errors for estimators for increasing strength of treatment effect modifiers at various sizes of trial-to-target difference. Bias was assessed as the difference in average treatment effect between trial estimates and the target population in the metric of standardized difference scores (*d*). Trial-to-target differences were assessed as odds ratios (OR)
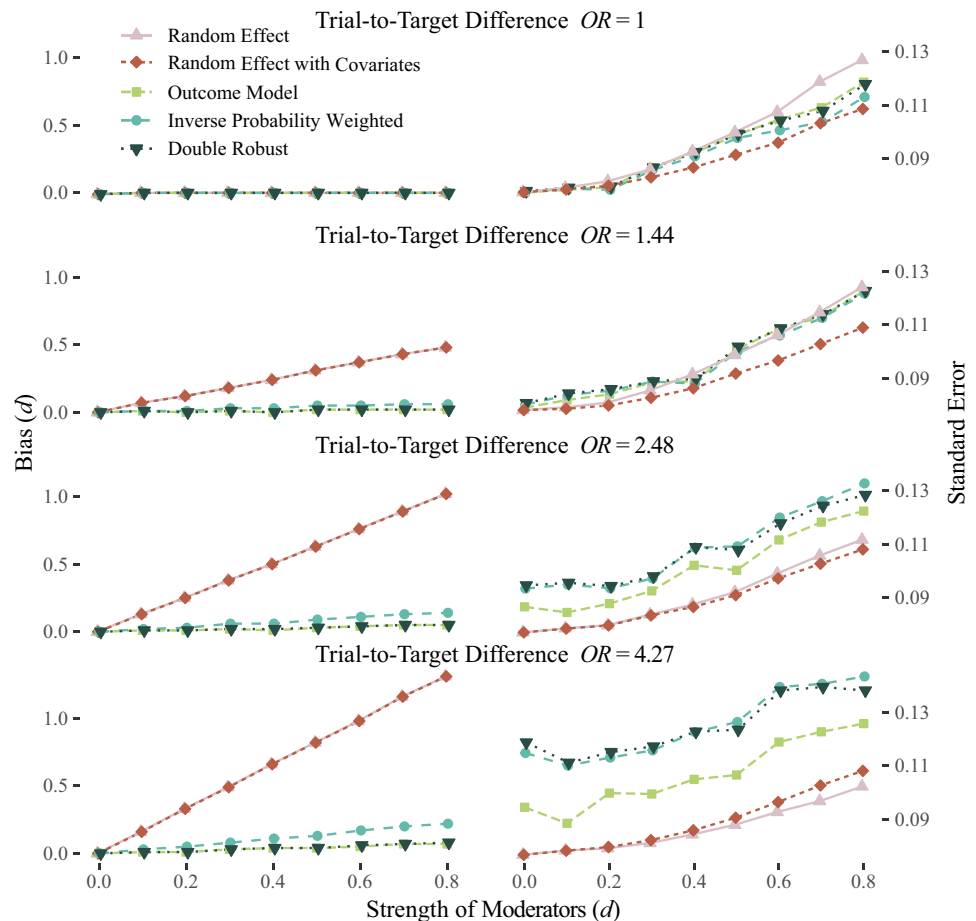
Fig. 2 Bias with increasing trial-to-target differences. Figure depicts bias and standard errors for estimators for increasing size of trial-to-target differences for various strengths of treatment effect modifiers. Strength of treatment effect modifiers was assessed as standardized difference scores (*d*). Trial-to-target differences were assessed as odds ratios (OR)
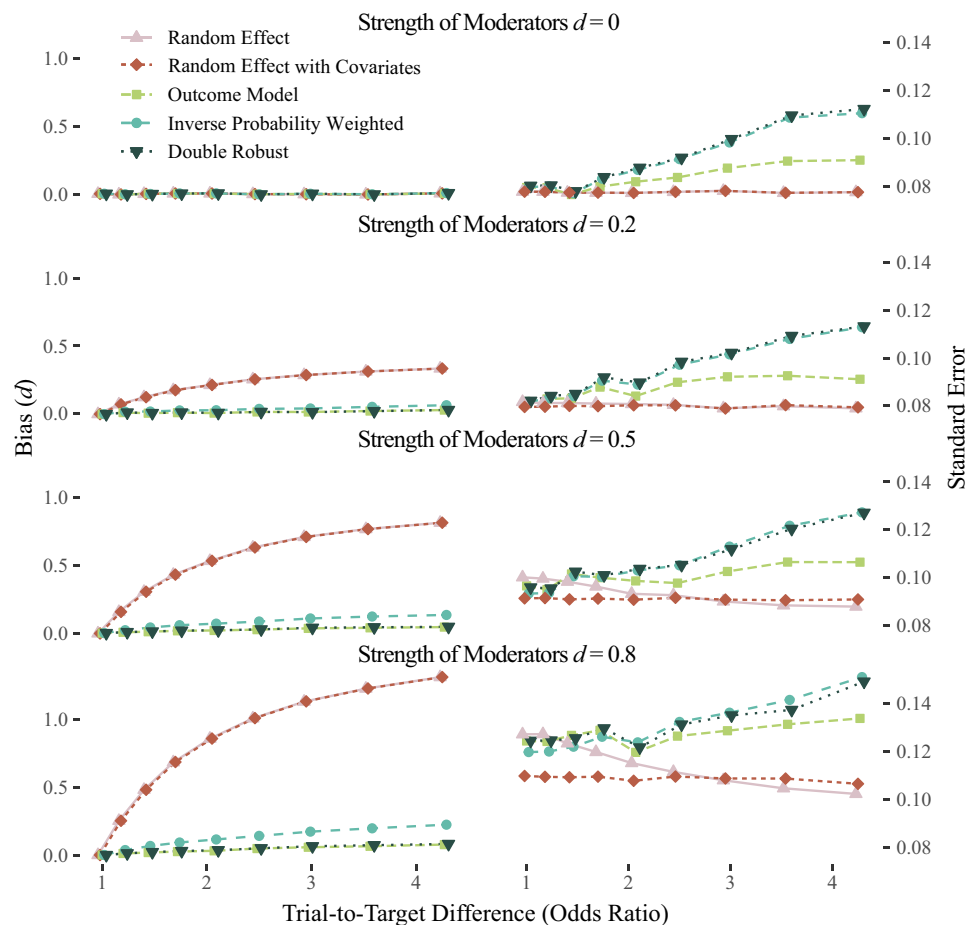
standard error as the number of trials with missing data increased from 0 to 60%. The stability of this estimate across amount of missingness suggests that the estimator produced an unbiased estimate of the treatment effect in the population defined by the pooled trials sample. However, this population is poorly defined as it is a mixture of two subpopulations that both differed from the target population. The increasing standard error was expected given increased uncertainty with increasing amount of missingness.

The pattern of bias in the DR estimators illustrates the impact of the additional missing at random assumption on bias of the estimators. Bias was lower in analyses that better met the missing at random assumption by including covariates with stronger associations to missing moderators. For the DR estimator with only weaker covariates, bias ranged from 0.05 to 0.25 when both moderators were missing. Bias was lower when stronger covariates were included (0.05 to 0.17) or when only one of the two moderators were missing (0.03 to 0.08). Standard errors for these approaches were similar across missing data patterns and higher than those from the multiply imputed estimates, although this difference diminished as the amount of missingness increased.

Bias for the DR estimator that only included trials with complete data showed no bias as the amount of missing data increased. This is not surprising given that the identifiability conditions were evaluated trial-by-trial and were met for trials with no missing data. However, the standard error for this approach increased markedly from 0.11 to 0.16 as missingness increased, reflecting the diminishing amount of information included in analyses as trials were excluded. These findings suggest an important trade-off between bias and uncertainty when considering which trials to include in analysis.

Discussion

Systematic missing data is present in most meta-analytic projects involving individual participant data. Bias is introduced when treatment effect modifiers are not appropriately accounted when estimating the treatment effect in the target population. Previous work addressing systematic missing data in meta-analysis with individual participant data have used the pooled sample across trials, which presumes all trials were drawn from the same target population (Burgess

Fig. 3 Bias with increasing systematic missing data. Figure depicts bias and standard errors for estimators for increasing amount of systematic missing data with trial-to-target difference set at an odds ratio (OR) of 2.07 and strength of the treatment effect modifiers set at a standardized difference score (*d*) of 0.4. There are three patterns of systematic missing data: (1) both modifiers were set to missing, (2) only one of the two modifiers was set to missing, and (3) two non-modifiers were set to missing
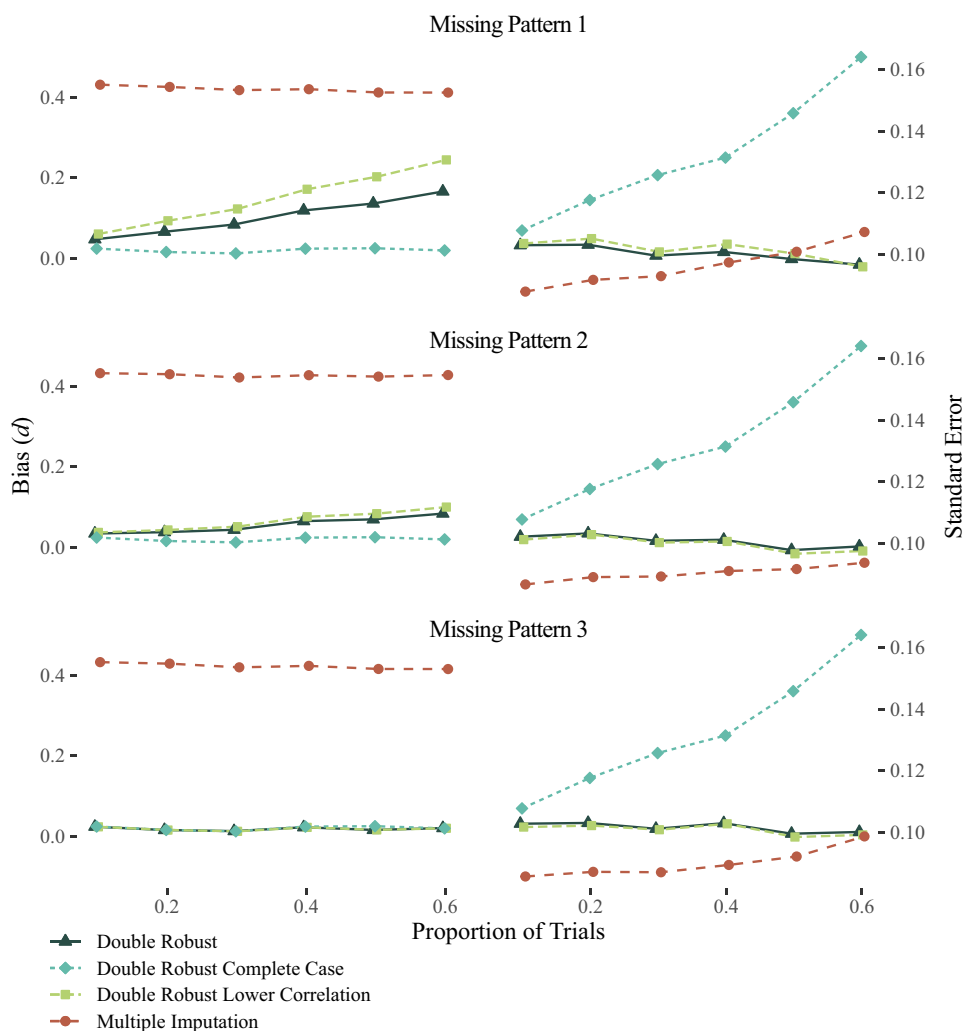
et al., 2013; Kunkel & Kaizar, 2017; Resche-Rigon et al., 2013; Resche-Rigon & White, 2018; Siddique et al., 2015, 2018). In practice, trial samples are almost always drawn from different populations that also differ from the target population suggesting that bias introduced from the mismatch between populations underlying the trials, and the target population is an important source of bias deserving careful consideration (Editors, 2021; Stuart et al., 2015; Susukida et al., 2016).

Unbiased Estimation in Target Population

Unbiased estimation of treatment effects in the target population requires five identifiability conditions. The first three are standard assumptions made when analyzing data from randomized trials and can hold for non-randomized designs when measured confounders are included in analysis. The last two conditions pertain to transporting causal inference from trials to a target population. The trials-to-target exchangeability assumption is conditional on covariates, meaning that it can be satisfied when all prominent treatment effect modifiers are identified, assessed, and accounted for in the analysis. The positivity of trail participation assumption suggests that every covariate pattern in the target population can occur in the trial samples. Unbiased estimation requires either that the trial data does not differ from the target in terms of prominent modifiers or that modifiers are measured in trial and target samples and adjusted for in the analysis. Simulations suggest that bias increases with larger trial-to-target differences in the distribution of treatment modifiers as well as when the strength of modifiers increases. Bias was present even with small trial-to-target differences and weak modification and increased as both dimensions increased. The amount of bias was striking and suggests that considering the match between trial evidence and target population requires close attention. Notably, the causal estimators that have been built for causal inference in meta-analysis (Steingrimsson et al., 2023) were unbiased in all full-data simulations.

Unbiased Estimation with Systematic Missing Data

Systematic missing data complicates the application of transportability methods as missing covariates cannot be used in modeling. Unbiased estimation is still possible if the covariate is not essential (e.g., not a treatment effect modifier) or if the distribution of the missing covariate is the same in both the trial and target populations. Bias introduced by including trials with systematic missing data on key covariates that are not from the same population as the target is important to consider. Simulations suggest that it may not be desirable to include all available evidence but rather all available evidence that can be accurately transported to the target population. The identifiability conditions described in this manuscript provide researchers with guidelines for deciding whether a trial should be included in analysis to provide unbiased estimates in the target population. Determining whether unbiased estimation is possible requires evaluating the identifiability conditions for each trial, or each set of trials that share the same pattern of systematic missing data.

The trial-by-trial evaluation of identifiability assumptions enables researchers to maximize the available information to make unbiased inference in the target population. For example, some trials may not have measured a treatment effect modifier but are expected to have a similar distribution of that modifier as the target population. Evidence from such trials could be included in analysis without adjustment. Other trials may differ from the target but have measured the modifier enabling inclusion through proper adjustment. Other trials may be expected to differ from the target and not have measured the effect modifier. It may be best not to include evidence from such trials as doing so would induce bias. The potential advantage of not including a trial with unmeasured treatment effect modification is seen in Fig. 2 as the causal estimator with no systematic missing data remained unbiased as the number of trials with missing data increased. However, this unbiased estimation came at the cost of precision, which is expected as the amount of usable data decreased with decreasing number of trials included in analysis. Decisions on the treatment of evidence from each trial requires considering the strength of the modifier, the difference in how the modifier is distributed in the trial compared the target, whether the modifier was measured in both the trial and target, and whether there are other covariates that could help meet the missing at random assumption for systematic missing covariates. Notably, these decisions shift with the consideration of each new target population.

Previous methods for systematic missing data using individual participant data rely on the pooled sample across trials. The pooled sample is a mixture of the underlying trials weighted by trial sample size and can be seen as a sample from the target population only when all trials sample from the same target population. In situations where all trials sample from the same target population conventional approaches are expected to produce unbiased estimation for that target, particularly when sampling variability is accounted for by addressing nesting of participants within trials (Burgess et al., 2013; Kunkel & Kaizar, 2017; Resche-Rigon et al., 2013; Resche-Rigon & White, 2018; Siddique et al., 2015, 2018). However, the pooled sample becomes problematic as trials are sampled from different populations because the weighted mixture no longer represents a real-world population. The pooled sample is also problematic when the population(s) underlying trial samples differs from the target population. Empirical evaluating the extent of differences among trial samples and the target requires a representative sample from the target population that assessed the same key covariates as the trials. In the absence of empirical evaluation, justification for pooled analyses requires theoretical arguments that all populations sampled by the trials are equivalent on key covariates such as treatment effect modifiers.

Missing Data in the Target Population

Defining the target population is key to making appropriate inference from trials. There has been considerable attention in guidelines and reporting standards given to the quality of evidence from the trials (Stewart et al., 2015), with less emphasis on the match between the evidence and the target population. The target population can be defined depending on the need of interested parties and can be defined at the level of a country, region, hospital, or program. When the target changes, so do the relative importance of the evidence from the trials, with those that are like the target increasing in importance. Evidence from trials that differ from the target either require less consideration or require models to support transportation of inference to the target. Transportability methods require a sample from the target population that measured the same treatment modifiers as the trials. There has been work showing that multiple imputation can be used for incomplete data in the target on the level of the participant, where the full data is available for some but not all participants (Hong et al., 2018). However, there are currently no available methods to address systematic missing data in the target. Given that the target is central to determining which trials can provide unbiased data and which ones can be transported, it is an important practical challenge to derive methods that allow for systematic missing data in the sample from the target population.

It can be difficult to find individual participant data from target populations that assessed key covariates using the same measures as the trials. The availability of samples from target populations depends on the area of investigation and on the focus of analysis. For example, research

in hospital systems can use electronic medical records to provide a sample of the target population. This presumes that the records contain key covariates such as treatment modifiers. We expect that population samples will become more accessible as electronic medical records, populations surveys, and other data sources work toward using common data elements and become more available to researchers (Chaimani, 2020; Zarin & Tse, 2016).

Identifying Treatment Modifiers

Identifying treatment effect modifiers is essential to the estimation of unbiased treatment effects in target populations. It is thus important to continue efforts to identify sources of heterogeneity of treatment effects (Lamont et al., 2018; Steingrimsson & Yang, 2019). Identifying heterogeneity has long been a goal in identifying appropriate treatments for subpopulations. Articulating the approaches and challenges of these efforts is beyond the scope of this manuscript (Dahabreh et al., 2016; Hu, 2023; Kent et al., 2020; Lagakos, 2006; Wang & Ware, 2011). It is important that potential modifiers be consistently reported in the literature. Given that most treatment trials are not powered to identify treatment modifiers, we suggest reporting effect sizes for the most influential modifiers even if they do not reach statistical significance (Lesko et al., 2018). Identifying prominent modifiers will help narrow the focus of which common data elements to include in trials and target samples and which variables to consider when evaluating the identifiability conditions.

Transporting Potential Outcome Means

In this manuscript, we have focused on transporting average treatment effects. Transporting potential outcome means offers additional benefits including comparing interventions that have not been compared in a head-to-head trial and including evidence from open trials (Barker et al., 2021). Unbiased transportation of potential outcome means requires stronger identifiability assumptions than for transportation of treatment effects, satisfying these assumptions requires identifying, assessing, and accounting for treatment effect modifiers and prominent predictors of the outcome (i.e., prognostic indicators) that differ between the populations underlying the trials and the target population (Dahabreh et al., 2020a, b). The annotated syntax included as supplementary materials first estimates the potential outcome means for each trial and can thus be generalized to situations that have different treatment conditions.

Conclusions

The formal identifiability conditions required to transport causal inference from trial samples to a target population highlight important considerations for evidence synthesis. The first and most important is recognizing bias introduced from mismatches between populations underlying the trials and the target population. Our simulations suggest that this bias can be considerable and evidence synthesis would benefit both from evaluation evidence quality from trials and careful evaluation of the match of that evidence to the target population. Specifically, it is important to consider the match on key covariates such as treatment effect modifiers or prognostic indicators. Second, it is important to have a sample from a well-defined target population that assessed key covariates using the same measures as the trials. Identifying appropriate samples is a barrier to both the evaluation and amelioration of bias due to a mismatch between evidence and the target population. Overcoming this barrier will require building on current efforts to use common data elements across trials and to make individual participant data publicly available (Ohmann et al., 2017; Polanin & Williams, 2016; Sheehan et al., 2016; Ventresca et al., 2020) and expanding those efforts to include key target populations. Third, while waiting for improved data infrastructure, investigators can maximize the use of available evidence by considering the match of each trial to the target and carefully considering which trials to include in analysis. Finally, our work highlights the importance of identifying key covariates for analysis. It is thus important to continue efforts to identify sources of heterogeneity of treatment effects (Lamont et al., 2018; Steingrimsson & Yang, 2019) by routinely reporting the size of potential treatment effect modifiers from clinical trials.

Declarations

Ethical Approval This study did not use participant data.

Informed Consent This study did not use participant data.

Conflict of Interest All authors declare that he/she has no conflict of interest.

Transparency and Replication We provided the R syntax for all simulations reported in this manuscript as supplemental material.

References

Barker, D. H., Dahabreh, I. J., Steingrimsson, J. A., Houck, C., Donenberg, G., DiClemente, R., & Brown, L. K. (2021). Causally interpretable meta-analysis: Application in adolescent HIV prevention. *Prevention Science*. https://doi.org/10.1007/s11121-021-01270-3

Brincks, A., Montag, S., Howe, G. W., Huang, S., Siddique, J., Ahn, S., Sandler, I. N., Pantin, H., & Brown, C. H. (2018). Addressing methodologic challenges and minimizing threats to validity in synthesizing findings from individual level data across longitudinal randomized trials. *Prevention Science, 19*(Suppl 1), 60–73. https://doi.org/10.1007/s11121-017-0769-1

Burgess, S., White, I. R., Resche-Rigon, M., & Wood, A. M. (2013). Combining multiple imputation and meta-analysis with individual participant data. *Statistics in Medicine, 32*(26), 4499–4514. https://doi.org/10.1002/sim.5844

Chaimani, A. (2020). Conduct and reporting of individual participant data network meta-analyses need improvement. *BMC Medicine, 18*(1), 156. https://doi.org/10.1186/s12916-020-01630-w

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., & Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal, 21*(1), C1–C68. https://doi.org/10.1111/ectj.12097

Dahabreh, I. J., Hayward, R., & Kent, D. M. (2016). Using group data to treat individuals: Understanding heterogeneous treatment effects in the age of precision medicine and patient-centered evidence. *International Journal of Epidemiology, 45*(6), 2184–2193. https://doi.org/10.1093/ije/dyw125

Dahabreh, I. J., Petito, L. C., Robertson, S. E., Hernán, M. A., & Steingrimsson, J. A. (2020a). Toward causally interpretable meta-analysis: Transporting inferences from multiple randomized trials to a new target population. *Epidemiology, 34*(8), 719–722. https://doi.org/10.1007/s10654-019-00533-2

Dahabreh, I. J., Robertson, S. E., Steingrimsson, J. A., Stuart, E. A., & Hernán, M. A. (2020b). Extending inferences from a randomized trial to a new target population. *Statistics in Medicine, 39*(14), 1999–2014. https://doi.org/10.1002/sim.8426

Dahabreh, I. J., Robertson, S. E., Tchetgen, E. J., Stuart, E. A., & Hernán, M. A. (2019a). Generalizing causal inferences from individuals in randomized trials to all trial-eligible individuals. *Biometrics, 75*(2), 685–694. https://doi.org/10.1111/biom.13009

Dahabreh, I. J., Robins, J. M., Haneuse, S. J.-P. A., & Hernán, M. A. (2019b). Generalizing causal inferences from randomized trials: Counterfactual and graphical identification. *[Stat]*. https://doi.org/10.48550/arXiv.1906.10792

Dahabreh, I. J., Steingrimsson, J. A., Robertson, S. E., Petito, L. C., & Hernán, M. A. (2019c). Efficient and robust methods for causally interpretable meta-analysis: Transporting inferences from multiple randomized trials to a target population. *[Stat]*. https://doi.org/10.48550/arXiv.1908.09230

Editors. (2021). Striving for diversity in research studies. *New England Journal of Medicine, 385*(15), 1429–1430. https://doi.org/10.1056/NEJMe2114651

Hernán, M. A., & VanderWeele, T. J. (2011). Compound treatments and transportability of causal inference. *Epidemiology, 22*(3), 368–377. https://doi.org/10.1097/EDE.0b013e3182109296

Hong, J.-L., Jonsson Funk, M., LoCasale, R., Dempster, S. E., Cole, S. R., Webster-Clark, M., Edwards, J. K., & Stürmer, T. (2018). Generalizing randomized clinical trial results: Implementation and challenges related to missing data in the target population. *American Journal of Epidemiology, 187*(4), 817–827. https://doi.org/10.1093/aje/kwx287

Hu, A. (2023). Heterogeneous treatment effects analysis for social scientists: A review. *Social Science Research, 109*, 102810. https://doi.org/10.1016/j.ssresearch.2022.102810

Jolani, S., Debray, T. P. A., Koffijberg, H., van Buuren, S., & Moons, K. G. M. (2015). Imputation of systematically missing predictors in an individual participant data meta-analysis: A generalized approach using MICE. *Statistics in Medicine, 34*(11), 1841–1863. https://doi.org/10.1002/sim.6451

Kent, D. M., Paulus, J. K., van Klaveren, D., D'Agostino, R., Goodman, S., Hayward, R., Ioannidis, J. P. A., Patrick-Lake, B., Morton, S., Pencina, M., Raman, G., Ross, J. S., Selker, H. P., Varadhan, R., Vickers, A., Wong, J. B., & Steyerberg, E. W. (2020). The predictive approaches to treatment effect heterogeneity (PATH) statement. *Annals of Internal Medicine, 172*(1), 35–45. https://doi.org/10.7326/M18-3667

Kunkel, D., & Kaizar, E. E. (2017). A comparison of existing methods for multiple imputation in individual participant data meta-analysis. *Statistics in Medicine, 36*(22), 3507–3532. https://doi.org/10.1002/sim.7388

Lagakos, S. W. (2006). The challenge of subgroup analyses — Reporting without distorting. *New England Journal of Medicine, 354*(16), 1667–1669. https://doi.org/10.1056/NEJMp068070

Lamont, A., Lyons, M. D., Jaki, T., Stuart, E., Feaster, D. J., Tharmaratnam, K., Oberski, D., Ishwaran, H., Wilson, D. K., & Van Horn, M. L. (2018). Identification of predicted individual treatment effects in randomized clinical trials. *Statistical Methods in Medical Research, 27*(1), 142–157. https://doi.org/10.1177/0962280215623981

Lesko, C. R., Henderson, N. C., & Varadhan, R. (2018). Considerations when assessing heterogeneity of treatment effect in patient-centered outcomes research. *Journal of Clinical Epidemiology, 100*, 22–31. https://doi.org/10.1016/j.jclinepi.2018.04.005

Markozannes, G., Vourli, G., & Ntzani, E. (2021). A survey of methodologies on causal inference methods in meta-analyses of randomized controlled trials. *Systematic Reviews, 10*(1), 170. https://doi.org/10.1186/s13643-021-01726-1

Ohmann, C., Banzi, R., Canham, S., Battaglia, S., Matei, M., Ariyo, C., Becnel, L., Bierer, B., Bowers, S., Clivio, L., Dias, M., Druml, C., Faure, H., Fenner, M., Galvez, J., Ghersi, D., Gluud, C., Groves, T., Houston, P., & Demotes-Mainard, J. (2017). Sharing and reuse of individual participant data from clinical trials: Principles and recommendations. *BMJ Open*, *7*(12). https://doi.org/10.1136/bmjopen-2017-018647

Polanin, J. R., & Williams, R. T. (2016). Overcoming obstacles in obtaining individual participant data for meta-analysis. *Research Synthesis Methods, 7*(3), 333–341. https://doi.org/10.1002/jrsm.1208

Resche-Rigon, M., & White, I. R. (2018). Multiple imputation by chained equations for systematically and sporadically missing multilevel data. *Statistical Methods in Medical Research, 27*(6), 1634–1649. https://doi.org/10.1177/0962280216666564

Resche-Rigon, M., White, I. R., Bartlett, J. W., Peters, S. A. E., Thompson, S. G., & PROG-IMT Study Group. (2013). Multiple imputation for handling systematically missing confounders in meta-analysis of individual participant data. *Statistics in Medicine, 32*(28), 4890–4905. https://doi.org/10.1002/sim.5894

Sheehan, J., Hirschfeld, S., Foster, E., Ghitza, U., Goetz, K., Karpinski, J., Lang, L., Moser, R. P., Odenkirchen, J., Reeves, D., Rubinstein, Y., Werner, E., & Huerta, M. (2016). Improving the value of clinical research through the use of Common Data Elements (CDEs). *Clinical Trials, 13*(6), 671–676. https://doi.org/10.1177/1740774516653238

Siddique, J., de Chavez, P. J., Howe, G., Cruden, G., & Brown, C. H. (2018). Limitations in using multiple imputation to harmonize individual participant data for meta-analysis. *Prevention Science, 19*(Suppl 1), 95–108. https://doi.org/10.1007/s11121-017-0760-x

Siddique, J., Reiter, J. P., Brincks, A., Gibbons, R. D., Crespi, C. M., & Brown, C. H. (2015). Multiple imputation for harmonizing longitudinal non-commensurate measures in individual participant data meta-analysis. *Statistics in Medicine, 34*(26), 3399–3414. https://doi.org/10.1002/sim.6562

Steingrimsson, J. A., Barker, D. H., Bie, R., & Dahabreh, I. J. (2023). Systematically missing data in causally interpretable meta-analysis. Biostatistics (Oxford, England), kxad006. https://doi.org/10.1093/biostatistics/kxad006

Steingrimsson, J. A., & Yang, J. (2019). Subgroup identification using covariate-adjusted interaction trees. *Statistics in Medicine, 38*(21), 3974–3984.

Stewart, L. A., Clarke, M., Rovers, M., Riley, R. D., Simmonds, M., Stewart, G., & Tierney, J. F. (2015). Preferred reporting items for a systematic review and meta-analysis of individual participant data: The PRISMA-IPD statement. *JAMA, 313*(16), 1657–1665. https://doi.org/10.1001/jama.2015.3656

Stuart, E. A., Bradshaw, C. P., & Leaf, P. J. (2015). Assessing the generalizability of randomized trial results to target populations. *Prevention Science, 16*(3), 475–485. https://doi.org/10.1007/s11121-014-0513-z

Stuart, E. A., Cole, S. R., Bradshaw, C. P., & Leaf, P. J. (2001). The use of propensity scores to assess the generalizability of results from randomized trials. *Journal of the Royal Statistical Society, 174*(2), 369–386. https://doi.org/10.1111/j.1467-985X.2010.00673.x

Susukida, R., Crum, R. M., Ebnesajjad, C., Stuart, E. A., & Mojtabai, R. (2017). Generalizability of findings from randomized controlled trials: Application to the National Institute of Drug Abuse Clinical Trials Network. *Addiction, 112*(7), 1210–1219. https://doi.org/10.1111/add.13789

Susukida, R., Crum, R. M., Stuart, E. A., Ebnesajjad, C., & Mojtabai, R. (2016). Assessing sample representativeness in randomized controlled trials: Application to the National Institute of Drug Abuse Clinical Trials Network. *Addiction, 111*(7), 1226–1234. https://doi.org/10.1111/add.13327

VanderWeele, T. J. (2009). Concerning the consistency assumption in causal inference. *Epidemiology, 20*(6), 880–883. https://doi.org/10.1097/EDE.0b013e3181bd5638

Ventresca, M., Schünemann, H. J., Macbeth, F., Clarke, M., Thabane, L., Griffiths, G., Noble, S., Garcia, D., Marcucci, M., Iorio, A., Zhou, Q., Crowther, M., Akl, E. A., Lyman, G. H., Gloy, V., DiNisio, M., & Briel, M. (2020). Obtaining and managing data sets for individual participant data meta-analysis: Scoping review and practical guide. *BMC Medical Research Methodology, 20*(1), 113. https://doi.org/10.1186/s12874-020-00964-6

Wang, R., & Ware, J. H. (2011). Detecting moderator effects using subgroup analyses. *Prevention Science, 14*(2), 111–120. https://doi.org/10.1007/s11121-011-0221-x

Zarin, D. A., & Tse, T. (2016). Sharing individual participant data (IPD) within the context of the Trial Reporting System (TRS). *PLOS Medicine*, *13*(1), e1001946. https://doi.org/10.1371/journal.pmed.1001946