American Journal of Epidemiology **Submitted Manuscript**

Title: Target Trial Emulation and Bias Through Missing Eligibility Data: An Application to a Study of Palivizumab for the Prevention of Hospitalization due to Infant Respiratory Illness.

Authors: Daniel Tompsett, Ania Zylbersztejn,   Pia Hardelid and Bianca De Stavola

Correspondence Address: Correspondence to Dr Daniel Tompsett, Population Policy and Practice Department, UCL GOS Institute of Child Health, United Kingdom, 30 Guilford Street, London, WC1N 1EH, email: d.tompsett@ucl.ac.uk

Affiliations: Population Policy and Practice Department, UCL GOS Institute of Child Health, United Kingdom, 30 Guilford Street, London, WC1N 1EH

Funding: MRC Methodology Grant, grant number: MR/R025215/1, which supported Daniel Tompsett and Bianca de Stavola, and a Wellcome Trust Seed Award in Science, grant reference number 207673/Z/17/Z, which supported Ania Zylbersztejn and Pia Hardelid

Data Availability Statement: The HTI is maintained by IQVIA (https://www.iqvia.com/), Copyright 2019, re-used with the permission of The Health and Social Care Information Centre. All rights reserved. Copyright 2021, re-used with the permission of IQVIA. All rights reserved. The authors do not have permission to share patient-level HTI data. Qualified researchers can request access to the data from IQVIA (contact Tanith Hjelmbjerg, tanith.hjelmbjerg@iqvia.com).

© The Author(s) 2022. Published by Oxford University Press on behalf of the Johns Hopkins Bloomberg School of Public Health. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

ORIGINAL UNEDITED MANUSCRIPT

Downloaded from https://academic.oup.com/aje/advance-article/doi/10.1093/aje/kwac202/6887820 by guest on 03 January 2023

Thanks: We would like to thank Hassy Dattani and Tanith Hjelmbjerg for their help with accessing the HTI database. This work uses data provided by patients and collected by the NHS as part of their care and support.

Conference presentation: This has not been presented as part of a conference presentation as of this time.

Preprint Information: NA

Disclaimer: None

Conflict of Interest: No conflicts of interest

Running Head: None

Key words: Target Trial Emulation, Eligibility, Missing Data, Multiple Imputation, Average Causal Effect

Abbreviations: MCAR: Missing Completely at Random, MAR: Missing at Random, MNAR: Missing not at Random, TTE: Target trial emulation, ACE: Average Causal Effect, G-age: Gestational Age, CI: Confidence Interval, RMSE: Root Mean Square Error, MCE: Monte Carlo Error.

Abstract

Target trial emulation (TTE) applies the principles of randomised controlled trials to the causal analysis of observational datasets. On challenge that is rarely considered in TTE is the sources of bias that may arise if the variables involved in the definition of eligibility into the trial are missing. We highlight patterns of bias that might arise when estimating the causal effect of a point exposure when restricting the target trial (TT) to individuals with complete eligibility data. Simulations consider realistic scenarios where the variables affecting eligibility modify the causal effect of the exposure and are Missing at Random (MAR) or Missing Not at Random (MNAR). We discuss multiple means to address these patterns of bias, namely, (i) controlling for the collider bias induced by the missing dataon eligibility, and (ii) imputing the missing values of the eligibility variables prior to selection into the TT. Results are compared to when TTE is performed ignoring the impact of missing eligibility. A study of Palivizumab, a monoclonal antibody recommended for the prevention of respiratory hospital admissions due

ORIGINAL UNEDITED MANUSCRIPT

Downloaded from https://academic.oup.com/aje/advance-article/doi/10.1093/aje/kwac202/6887820 by guest on 03 January 2023

to Respiratory Synctial Virus in high risk infants, is used for illustrations.

Keywords: Target Trial Emulation, Eligibility, Missing Data, Multiple Imputation, Average Causal Effect

1. Introduction

Randomised controlled trials (RCTs) are commonly used for estimating causal effects of point interventions. However in many epidemiological settings, an RCT may be infeasible or ethically nonviable. Hence, observational data are also used to compare effectiveness, with various strategies adopted to address the lack of randomization, and indication bias, for example by controlling for measured confounders. Analysis of observational data suffers from various additional sources of bias such as selection bias, indication bias, and immortal time bias [1].

Target trial emulation (TTE) aims to avoid some of these biases by adopting the design principles of RCTs. Individuals in an observational database, such as administrative health records, are selected according to a set of eligibility criteria that mirrorthose that would be used in an RCT [2]. However, data on variables that determine eligibility are often incomplete, and as such not all participants of the Target Trial (TT) are identifiable from the observational database. It is typically advised to consider a different target trial with more complete eligibility criteria ([1]), or to exclude or censor individuals with missing data [3,4]. Missing data is often a source of bias when those excluded are systematically different from the observed, *i.e.* if they are missing at random (MAR) or missing not at random (MNAR) [5,6]. Though identified as a potential limitation, there is little work investigating the extent to which missing eligible data can impact the analysis of a target trial.

One solution is to impute missing eligibility prior to selection into a target trial. However, we could find only one precedent of imputation of eligibility criteria prior to the creation of a target trial in [7]. More generally, multiple imputation of exclusion criteria in observational studies has been considered in a recent work [8] for validating error prone confounders, but remains an infrequently studied topic. We intend to bring attention to work of this kind to the context to TTE.

In this paper we investigate biases in the average causal effect (ACE) of a point exposure, in a target trial with missing eligibility data. Our simulations consider realistic scenarios where the eligibility variables modify the true causal effect. We consider two strategies of analysis: (i) Conditioning on variables that drive missingness eligibility (ii) recovering the missing eligibility data via multiple imputation (MI). A study of Palivizumab, a monoclonal antibody for prevention of symptoms of severe Respiratory Synctial Virus (RSV) infection in high risk infants based on administrative hospital and pharmacy dispensing data is used to illustrate these alternative approaches.

2. Methods

1. Setup

Consider the setting with a binary treatment A, end of study outcome Y, and confounding variables L_1 and E, where the latter determines eligibility. Suppose E has informative missingness with R_E an indicator of completeness (1=complete, 0=missing). Missingness in E may be missing at random (MAR), driven by variables that are not necessarily confounders, which we denote L_2 and L_3, or Missing not at Random (MNAR), if also driven by E itself [9] (Figure 1). This is a typical setting, whereby L_2 and L_3 are separate causes of respectively A and Y.

ORIGINAL UNEDITED MANUSCRIPT

Downloaded from https://academic.oup.com/aje/advance-article/doi/10.1093/aje/kwac202/6887820 by guest on 03 January 2023

We emulate a TT where eligibility is defined by E being greater than or equal to some value e. In practice E may represent a set of variables, which determine an eligibility indicator variable I_E. The mechanism for inclusion is shown in Figure effg:dag2, represented by the box (indicating conditioning) surrounding $I_E = 1$. It also shows the selection mechanism induced by restricting the TT to individual with complete E, indicated by the box around $R_E = 1$. We distinguish between:

- The source population, from which the TTs are derived.
- The full eligibility TT (TT_{true}), containing all those who are eligible $(I_E = 1)$.
- The complete eligibility TT (TT_{obs}), containing those who are complete and eligible, ($I_E = 1$ and $R_E = 1$).

Our target estimand is the average causal effect (ACE) of A on Y in TT_{true}, defined as,

$$\mathrm{ACE}^{I_E=1} = \mathbb{E}(Y(1) - Y(0)|I_E = 1), \tag{1}$$

where $E(Y(a))$ is the average value of Y, if the exposure A were set to take the value $a, for\ a = 0{,}1$ in the whole population. In reality, TT_{true} is not known, and thus $\mathrm{ACE}^{I_E=1}$ is approximated by the equivalent estimand from TT_{obs},

$$\mathrm{ACE}^{I_E=1,R_E=1} = \mathbb{E}(Y(1) - Y(0)|I_E = 1, R_E = 1). \tag{2}$$

The ACE of a point exposure can be identified by invoking assumptions of no interference, counterfactual consistency, and conditional exchangeability (*i.e.* no unmeasured confounding) [2].

2. *Sources of Bias*

If we attempt to estimate $ACE^{I_E=1}$ from an estimate of $ACE^{I_E=1,R_E=1}$ we would be prone to two sources of bias.

Collider bias

The confounders L_1 and E are common causes of exposure and outcome which need to be controlled for, whilst L_2 and L_3, the drivers of missingness, are not. However when we condition on $R_E = 1$, we create a spurious association between L_2 and L_3, which confounds the causal effect of A on Y via $A \to L_3 \to L_2 \to Y$ (Figure 2). This is a type of collider bias known as Berkson's Bias [10,11], which must be removed by conditioning on either L_2or L_3.

Selection bias

When E has informative missing data, the missing eligible ($I_E = 1, R_E = 0$) contain information about TT_{true} that cannot be recovered by TT_{obs}. This can result in selection bias when conducting an analysis on TT_{obs} if, for any reason, the causal effect of A on Y is different in the missing eligible, compared to the complete eligible.

By controlling for L_1, L_2, and E we identify the causal effect,

$$\begin{aligned}\mathrm{ACE}_{(L_1,L_2,E)}^{I_E=1,R_E=1} &= \mathbb{E}(Y(1) - Y(0)|L_1 = l_1, L_2 = l_2, E = e, I_E \\ &= 1, R_E = 1)\ \forall\ (l_1, l_2, e).\end{aligned} \tag{3}$$

To find $\mathrm{ACE}^{I_E=1,R_E=1}$ we marginalise (average) $\mathrm{ACE}_{(L_1,L_2,E)}^{I_E=1,R_E=1}$ over the distribution of L_1, L_2 and E in TT_{obs}.

If the effect of A on Y is modified by these confounders, then the value of $\mathrm{ACE}^{I_E=1,R_E=1}$ depends on the distribution of that confounder in TT_{obs}.

Hence since we cannot recover the distribution of the confounders in TT_{true}, $\mathrm{ACE}^{I_E=1,R_E=1}$, obtained from TT_{obs}, is a biased approximation of $\mathrm{ACE}^{I_E=1}$. In other words the distribution of the confounders in TT_{obs}, does not match that in TT_{true}.

ORIGINAL UNEDITED MANUSCRIPT

Downloaded from https://academic.oup.com/aje/advance-article/doi/10.1093/aje/kwac202/6887820 by guest on 03 January 2023

Suppose E was a score capturing standards of hospital care. We might expect a treatment A to be more effective on the outcome at higher standards of care. Now if hospitals of a low standard are more likely to have a missing score, then we would over-represent eligible hospitals of higher standards in TT_{obs}, and lead to a biased ACE.

Dealing with this bias requires recreating the joint distribution of exposure, outcome and confounders of TT_{true} for example using multiple imputation ([12,13,14]).

This bias has been discussed in the wider setting of "data-fusion" of multiple data sources ([15]), with identification of targeted causal effects involving knowledge of the distribution of the confounders in the "fused" population, as we discuss above. This type of bias has been referred to as an issue of "transportability" [15] or external validity to a different population [16]. **Our setting is created by missing information that precludes the identification of the target population. This could be viewed as an issue of internal validity of TT_{obs} itself, or of its external validity to TT_{true}. The issue also impacts the generalisability of results to other populations.**

3. Strategies

We indicate possible strategies to address the biases in the estimation of $\text{ACE}^{I_E=1}$ above.

Strategy 1: Ignoring missing eligibility

In the setting of Figure 2 we fit an outcome regression model for Y on A, controlling for L_1 and E in the model, and then estimate $\text{ACE}^{I_E=1,R_E=1}$ by marginalising over their distribution in TT_{obs}, as described in [17].

Strategy 2: Dealing with collider bias

With this approach we fit an outcome regression model for Y on A controlling for L_1, E and either L_2 or L_3 in order to block the path opened by conditioning on R_E, and then estimate $\text{ACE}^{I_E=1,R_E=1}$ as in strategy 1.

If the estimand of interest is $\text{ACE}^{I_E=1,R_E=1}$, then this strategy is sufficient to remove bias induced by missing eligibility data.

Strategy 3: Dealing with collider and selection bias

We specify an imputation model (IM) to predict the missing eligibility data in the source population. We impute E in multiple copies of the source population, and from each, construct an imputed copy of TT_{true} using imputed eligibility criteria. Wethen control for L_1 and E as in strategy 1 to estimate $\text{ACE}^{I_E=1}$ in each copy, which are pooled using Rubin's Rules [9].

Implementation

The imputation step is as follows:

1. Specify an IM for the missing mechanism of E.
2. Generate m copies of the source population and impute E in each copy based on the IM.
3. Apply the eligibility criteria to each imputed dataset to obtain m emulated versions of TT_{true}.
4. Estimate $\text{ACE}^{I_E=1}$ in each imputed TT_{true}, controlling for L_1 and E to obtain m estimates of the causal effect of A on Y, $\widehat{ACE}_m^{I_E=1}$.
5. Obtain Rubin's pooled estimate of the target causal effect by taking the average over the m imputed sets:

ORIGINAL UNEDITED MANUSCRIPT

Downloaded from https://academic.oup.com/aje/advance-article/doi/10.1093/aje/kwac202/6887820 by guest on 03 January 2023

$$\widehat{ACE}^{I_E=1} = \frac{1}{m}\sum_{i=1}^{m} \widehat{ACE}_m^{I_E=1}$$

6.

To capture any suspected treatment effect heterogeneity, imputations are carried out separately for each value of A. Note that this technique requires A be fully observed [18].

Work in [8] highlights that excluding data relevant to inclusion in a study after MI leads to biased estimates of Rubin's pooled estimate of the variance because of incongeniality between the imputation and outcome model. We hence consider confidence intervals using a percentile based bootstrap.

Combining Bootstrap and Imputations

We combine bootstrapping with MI using the "Boot-MI" methodology [19]. This consists of the following steps:

1. Obtain b bootstrap samples of the source population.
2. Apply steps 1-5 of the MI procedure above for each of the b datasets and obtain b estimates of $\widehat{ACE}_b^{I_E=1}$.
3. A percentile based bootstrapped confidence interval is then derived as the $\alpha \times 100^{th}$ and $(1-\alpha) \times 100^{th}$ percentiles of the ordered bootstrapped estimates.

We use single imputation $(m = 1)$ which has been shown to have good statistical properties [20], and reduce computational burden ([20,19]), nested within $b = 1000$ bootstraps, which is at or above the typically recommended number ([21]).

Sensitivity analyses

Imputation models for E that allow for different mean values depending on A could be used

$$\mathbb{E}(E|Y, A = a, L_1, L_2, L_3, R_E) = \beta_0 + \sum_{i=1}^{3} \beta_i L_i + \beta_5 Y + \delta_a R_E \text{ for } a = 0{,}1.$$

We use fully conditional specification (FCS or MICE) using the "`mice`" package in `R` [13,22] to impute the data. The parameters δ_a are MNAR sensitivity parameters. If MNAR is suspected, setting $\delta_a \neq 0$ shifts the imputedvalues of E (separately for each a) by an amount that accounts for the effect of E on its own missingness ([23,24]). In practice, sensible ranges for δ_a are chosen, with the data imputed over these ranges.

3. Simulations

We investigate strategies 1-3 by simulating data according to the structure of Figure 2. Specifically:

- L_1,L_2 and L_3 are independent $N(0,1)$.
- E is a normal variable dependent on L_1 and L_2:

$$E \sim N(L_1 + L_2, 1).$$

-
- Eligibility is defined as $I_E = 1$ if $E \geq 0$, $I_E = 0$ otherwise, hence around 50% of the population are eligible.
- The missing mechanism of E is expressed as a linear function of L_2, L_3, and E:

$$log(\text{odds}_{R_E}) = \mu + \alpha L_2 + \alpha L_3 + \gamma E.$$

-
- The exposure A is a binary variable, and generated in terms of the log-odds of exposure, expressed as a linear function of L_1, L_3, and E:

$$log(\text{odds}_A) = 0.1L_1 + 0.5L_3 + 0.1E.$$

ORIGINAL UNEDITED MANUSCRIPT

Downloaded from https://academic.oup.com/aje/advance-article/doi/10.1093/aje/kwac202/6887820 by guest on 03 January 2023

- Around 54% of individuals in the source population are exposed.
- The outcome Y is a normal variable that depends on exposure A, eligibility E, their interaction, and also on L_1 and L_2, with L_2 exercising a stronger impact than L_1:

$$Y \sim N(A + E + AE + L_1 + 2L_2, 1).$$

-

The source population is of size $n = 1{,}000$. We investigated strategies 1-3 at different values of μ, α and γ, the parameters affecting R_E. Specifically, μ drives the percentage of MCAR missingness. α drives the strength of the MAR assumption, and the spurious association between L_2 and L_3, and γ drives the strength of the MNAR mechanism, with positive values leading to a higher probability of larger values of E being observed.

The parameter μ was set at 0 and 1.5, leading to severe (50%) and moderate (18%) MCAR missingness. α and γ were set to range from 0 (no association) up to ± 0.4. For each combination we carried out $l = 1{,}000$ simulations for each of these scenarios using $b = 1000$ bootstraps, reporting for each the average bias in the estimation ($ACE^{I_E=1,R_E=1} - ACE^{I_E=1}$), its Monte Carlo Error (MCE), Root Mean Squared Error (RMSE) and 95% coverage [25].

4. Results

Observed and True Target Trial Comparisons

Table 1 describes the characteristics of a set of single large simulations of TT_{obs} for different values of α, γ and μ. We set $n = 1{,}000{,}000$ to minimise random variation. The three missingness scenarios are MCAR ($\alpha = \gamma = 0$),MAR ($\alpha \neq 0$ and $\gamma = 0$), and MNAR ($\alpha \neq 0$ and $\gamma \neq 0$). The scenario when E is not missing, (TT_{true}) is included for comparison.

When the mechanism is MCAR, the means and correlations of relevant variables are not affected. When the mechanism is MAR, they depart from those found in TT_{true}: when $\alpha > 0$, individuals in TT_{obs} have larger mean values for E, L_2 and L_3 than in TT_{true}. This is because α leads to individuals with larger values for L_2 and L_3 being more likely to be observed, shifting upwards their distributions, and by extension, the distribution of E. When α is negative the opposite is true. These biases are more noticeable at $\mu = 0$ due to the greater proportion of missing individuals.

Under MNAR, setting $\gamma > 0$ makes higher values of E more likely to be observed in TT_{obs}, with the opposite occuring when $\gamma < 0$, leading to shifts in the distributions for E, L_2 and L_3 similar to what occurs with α.

The combined impact of α and γ varies. When both are of the same sign, their impacts compound, and strengthen the corresponding shifts in distribution. When they are of opposite sign, their impacts partially offset one another.

The shifts in distribution for L_1 are complicated, shifted **downwards** when $\alpha > 0$ but shifted **upwards** when $\gamma > 0$. This is due to a complicated relationship between the spurious negative $L_1 - L_2$ association (caused by conditioning on I_E), driving a downward shift in L_1 with higher values of L_2 is, and the positive $L_1 - E$ association, driving an upward shift with higher values of E.

Strategies

For strategies 1 and 2, bias in estimation of ACE increased with higher values of alpha and gamma, and was worse when mu=0 (Tables 2 and 3). This is due to having to average over the distribution of the confounders to estimate $ACE^{I_E=1}$. The size and direction of this bias is

ORIGINAL UNEDITED MANUSCRIPT

Downloaded from https://academic.oup.com/aje/advance-article/doi/10.1093/aje/kwac202/6887820 by guest on 03 January 2023

nearly identical to the shift in the distribution of E observed in Table 1. This is because effect modification by E has effect size equal to 1.

The impact of collider bias induced by α is negligible, as shown by the small differences in bias for strategies 1 and 2. The RMSE is smaller for strategy 2, but has more under-coverage, possibly because it involves averaging L_2, which also has a shifted distribution.

Table 1 implies that had L_2 been the effect modifier rather than E, strategies 1 and 2 would have shown more bias under the MAR assumption. This is investigated, in Web Appendix 1.

Strategy 3 shows unbiased estimates (Within MCE) in all cases indicating a successful recovery of the causal effect in TT_{true}. The CIs however display over-coverage, particularly when a large fraction of the eligible are missing.

Selection bias appears to increase under the following conditions:

- Larger numbers of missing eligible individuals.
- Larger values of α and γ, the drivers of missingness.
- A stronger effect modification of the causal effect of A on Y by E (or any variables related to E).

With fewer eligible participants lost to missingness, there is less missing data to drive a differentiation in the distributions of E in TT_{obs} and TT_{true}, which is why bias decreased when μ was larger, and the number of missing eligible decreased. None of these features are likely to be known in advance.

When E was MNAR, imputation was carried out with the correct values of the sensitivity parameters δ_0, δ_1. This was to demonstrate that, all other biases (including a mis-specified imputation model) accounted for, strategy 3 can eliminate the biases of Section 2.2 when E is MNAR. This is unlikely to be possible in reality, hence in Web Appendix 2 we repeat specific MNAR simulations of Table 3 assuming a MAR imputation model $(\delta_0, \delta_1) = (0, 0)$, which shows notable bias. This highlights that in practice, MNAR imputation is an exploratory technique, and careful consideration must be taken to choose informative values of ranges for δ_0 and δ_1 to investigate ([23,24]). A realistic application of strategy 3 is shown in the case study.

In summary, strategy 3 is necessary in the case that missing data are noticeably MAR or MNAR. If not the case a user may prefer the simpler strategies 1 and 2. Strategy 2 is the most precise if this is preferred by the user, but one must account for the possibility of undercoverage if a CI is sought.

5. Case Study: Effect of Palivizumab on Infant Hospital Admission

Respiratory Syncytial Virus is a major cause of acute lower respiratory tract infection in infants, with RSV bronchiolitis responsible for 40,000 hospital admissions annually in England [26]. Palivizumab is licensed for passive immunisation to prevent RSV in premature infants with Congenital Heart Disease (CHD) or Chronic Lung Disease (CLD). Due to its high cost, Palivizumab is typically recommended to more select groups of high risk infants than those in clinical trials, with limited data on real world effectiveness [27]. Hence analysis by a selective emulated trial is of interest.

An observational cohort of infants potentially eligible for Palivizumab treatment in England has been developed ([27]), using the Hospital Treatment Insights (HTI) database, which links pharmacy dispensing records from 43 acute hospitals in England, and hospital records from Hospital Episode Statistics (HES). This cohort details infants born between 1st Jan 2010 and 31st December 2016, with follow up data on Palivizumab prescriptions and hospital admission up to their first year of life. HTI is maintained by IQVIA `https://www.iqvia.com/`.

ORIGINAL UNEDITED MANUSCRIPT

Downloaded from https://academic.oup.com/aje/advance-article/doi/10.1093/aje/kwac202/6887820 by guest on 03 January 2023

This cohort identifies a source population of 8294 high risk infants, defined as having CLD or CHD, under care of an HTI-reporting hospital, alive at the start of their first RSV season (October 1st to the 31st March), with a full linked hospital admission history. This is shown in the cohort flowchart of Figures 3 and 4.

Infants in the source population were considered eligible for the TT if they had a diagnosis of CLD or CHD and met additional eligibility criteria based on gestational age and chronological age at start of RSV season. Specifically, those who met criteria 1a or 2a in Chapter 27a for recommendation of treatment by Palivizumab in the Green Book [28] (Web Table 3). Gestational age however is missing for 2814 (34%) infants in the source population. As a result, the eligibility of many children cannot be identified.

Target Trial Emulation

The emulated target trial protocol is detailed in Web Appendix 3 and Web Table 2. We define TT_{obs} to include all eligible individuals with complete eligible data on gestational age, birthweight, index of multiple deprivation (IMD) score and ethnicity. This led to a trial of 1560 infants. We also aim to recover TT_{true} by imputing missing gestational age in the high risk cohort. This corresponds to using strategies 2 and 3 respectively.

We are interested in the effect of any Palivizumab prescription on RSV related hospital admission in infants during their first RSV season of life. A full course of treatment by Palivizumab requires up to five monthly doses during RSV season. As we could not determine adherence to treatment from the HTI data, we define a simplified exposure as a binary indicator of having been prescribed at least one dose of Palivizumab in their first RSV season of life. Infants are identified in the first month of life for treatment, and typically administered in outpatient clinics, not when hospitalised for RSV. Our outcome is a binary indicator of having been hospitalised for an RSV related condition during their first RSV season of life.

Our target estimand is the ACE of palivizumab prescription on RSV related hospital admission in TT_{true}, expressed as the average difference in absolute risk of hospital admission (the intent to treat (ITT) effect).

To balance the confounders in the treated and untreated, we fit a model for the propensity of receiving Palivizumab including gestational age, age at start of RSV season, IMD quintiles, sex, ethnicity, year of birth, diagnosis of CLD or CHD, or both, andother comorbidities. The resultant propensity scores showed reasonable overlap in the treated and untreated (Web Figure 1). Mean differences between treated and untreated, adjusted for inverse probability of weighting by propensity score, were within 0.1, indicating good confounder balance.

We fit two different outcome models, a logistic regression model of hospital admission against treatment with inverse probability weight of being treated (IPTW), corresponding to a Marginal Structural Model (MSM) [29], and a second where we control for the propensity score, and all confounders directly in the outcome model, similar to those in two stage g-estimation of Structural Nested Mean Models (SNMMs) [30]. The ACE is calculated by estimating potential outcomes via the "data stacking" method of [17].

Continuous gestational age is imputed in the treated and untreated arms separately (to account for any interaction between gestational age and Palivizumab) using a MNAR imputation model that includes all the variables of the propensity score model, plus the outcome and birthweight. Birthweight is not included in the outcome model due to collinearity with gestational age. There are thus two sensitivity parameters δ_1 for exposed and δ_0 for unexposed. We assert that infants with missing gestational age may have higher mortality, implying a shorter gestation [31]. Hence we run the analysis setting δ_1 and δ_0 to either 0 (MAR), or -4 (MNAR). Based on recommendations in [23], rather than compare the ACE directly to δ_0 and δ_1, which are difficult to interpret physically, we estimate from the imputed data the mean gestational age

ORIGINAL UNEDITED MANUSCRIPT

Downloaded from https://academic.oup.com/aje/advance-article/doi/10.1093/aje/kwac202/6887820 by guest on 03 January 2023

in treated and untreated infants to contrast against the results.

Missing birthweight, IMD score and ethnicity and imputed alongside gestational age using MICE. We report the results in Tables 4 and 5 below.

Results

Analysis of TT_{obs} suggests that treatment by at least one dose of Palivizumab has little effect on the risk of being hospitalised, indicated by an ACE of -0.003 using a propensity score conditioned outcome model, and -0.01 under IPW (a 0.3% or 1.0% lower risk of hospital admission). When imputing the TT under MAR we observe a 0.1% and 0.2% lower risk of hospital admission respectively. Under MNAR there is a more noted effect of Palivizumab, ranging from $-1.0 - 1.3\%$ using a outcome model controlled for the propensity score, and $-3.1 - 2.3\%$ using IPW.

The imputation model implies a high number of missing eligible participants, with over 1000 more individuals under MAR imputed trial, and up to nearly 2500 more under MNAR.

When δ_0 was set to -4, this led to a reduction in average gestational age in the untreated by 2.2 weeks. In this case there was stronger reduction in risk of hospital admission when treated. When δ_1 was set -4, the average gestational age in the treated was reduced by 2.7 weeks and there was an increasing risk of hospital admission under treatment.

No estimate was found to be significant based on a 95% CI. Despite there being a clear change in the distribution of gestational age under MNAR conditions, and a large number of missing eligible there is only weak evidence of selection bias in this study. This implies that gestational age only weakly modifies the effect of Palivizumab on hospital admission.

The implication is that receiving at least one dose of Palivizumab appears to have little effect on hospital admission, and are robust to changes in the missing data assumption.

6. Discussion

In this paper we bring to light notable sources of bias in TTE, emanating from ignoring missing eligible data. We explore one means to analyse a TT combined with multiple imputation of eligibility criteria prior to selection. We demonstrate via simulationthat an imputed TT can eliminate sources of selection and collider bias, improve the sample size of a TT and allow users to investigate sensitivity to changes in the assumptions of the missing eligible data on effect size.

An imputed TT for the effect of receiving at least one dose of Palivizumab on RSV related hospital admission indicated a significant number of infants with missing gestational age were eligible, though any selection bias in this case was small.

We identified characteristics of the data that determine the size of selection bias, namely the strength of the MAR or MNAR mechanism, the number of missing eligible individuals and the size of the effect modification. None of these characteristics can becalculated from the source population but could be inferred using external linked datasets. This selection bias can occur if any variable related to eligibility is an effect modifier. We show in Web Appendix 1, that when L_2 was the effect modifier, strong selection bias was identified when E was MAR.

A limitation of the method is the tendency of confidence intervals to over-cover. The Boot-MI method is computationally intensive and thus one should expect an analysis to take several hours even with, hence we construct CIs using a percentile bootstrap with just single imputation. However single imputation lends itself to overcoverage [19]. In Web Appendix 2 we apply strategy 3 using MI with $m = 5$, which demonstrates improved coverage. One alternative would be to investigate the corrected Rubin's pooled variance of $ACE^{I_E=1}$ suggested in [8]. However, obtaining accurate confidence interval estimates in this way for the

ORIGINAL UNEDITED MANUSCRIPT

Downloaded from https://academic.oup.com/aje/advance-article/doi/10.1093/aje/kwac202/6887820 by guest on 03 January 2023

ACE using MI requires complex methods [32,33,34].

Instead of MI, we could consider using Inverse Probability Weighting to address the bias caused by missingness in E ([35]). We investigate this method in Web Appendix 2, and found it did not correct the bias. Another possible alternative is toutilize the work in [15], by inferring or presuming the distribution of the confounders in TT_{true}, and standardising the conditional ACE estimated in TT_{obs}, but would be a considerable challenge.

It is also worth noting that using strategy 2, and targeting the causal effect in those with complete records may be a pragmatic choice if the expected selection bias is limited and the source population is cumbersome.

As data on Palivizumab prescriptions and adherence were limited, this impacted the quality of conclusions that could be made. Clinical colleagues reassure us that children hospitalised with RSV would not be issued palivizumab, protecting from reverse causation. However, other issues such as confounding by indication cannot be discounted. Limitations of the diagnostic data also meant a slight inflation of our definition of the eligible population because some of the diagnoses may include less severe diseases than listed in the Green Book.

Acknowledgments and Funding

All authors are affiliated with the UCL Great Ormond Street Institute of Child Health, which is supported by the NIHR Great Ormond Street Hospital Biomedical Research Centre. This work was supported by the MRC Methodology Grant, grant number: MR/R025215/1, which supported Daniel Tompsett and Bianca de Stavola, and a Wellcome Trust Seed Award in Science, grant reference number 207673/Z/17/Z, which supported Ania Zylbersztejn and Pia Hardelid.

With regards to data availability. The HTI is maintained by IQVIA (https://www.iqvia.com/), Copyright ©2019, re-used with the permission of The Health and Social Care Information Centre. All rights reserved. Copyright ©2021, re-used with the permission of IQVIA. All rights reserved. The authors do not have permission to share patient-level HTI data. Qualified researchers can request access to the data from IQVIA (contact Tanith Hjelmbjerg, tanith.hjelmbjerg@iqvia.com). All study protocols are subject to review by an Independent Scientific and Ethical Advisory Committee (ISEAC).

We would like to thank Hassy Dattani and Tanith Hjelmbjerg for their help with accessing the HTI database. This work uses data provided by patients and collected by the NHS as part of their care and support. Note that all analyses, and simulations were performed in `R` [22], and are available on request. For details see Web Appendix 4.

Conflicts of Interest

We declare no conflicts of interest.

References

[1] Hernán Miguel A., Robins James M.. Using Big Data to Emulate a Target Trial When a Randomized Trial Is Not Available *American Journal of Epidemiology.* 2016;183:758-764.

[2] Hernán M A. *Causal Inference: What If*. Boca Raton, Florida: Chapman and Hall CRC 2020.

[3] Kutcher Stephen A., Brophy James M., Banack Hailey R., Kaufman Jay S., Samuel Michelle. Emulating a Randomised Controlled Trial With Observational Data: An Introduction to the Target Trial Framework *Canadian Journal of Cardiology.* 2021.

[4] Maringe Camille, Benitez Majano Sara, Exarchakou Aimilia, et al. Reflection on modern

ORIGINAL UNEDITED MANUSCRIPT

Downloaded from https://academic.oup.com/aje/advance-article/doi/10.1093/aje/kwac202/6887820 by guest on 03 January 2023

methods: trial emulation in the presence of immortal-time bias. Assessing the benefit of major surgery for elderly lung cancer patients using observational data *International Journal of Epidemiology.* 2020;49:1719-1729.

[5] Little Roderick, Rubin. Donald. *Statistical Analysis with Missing Data, Third Edition*. Hoboken, NJ, USA: Wiley 2019.

[6] White Ian R., Kalaitzaki Eleftheria, Thompson Simon G.. Allowing for missing outcome data and incomplete uptake of randomised interventions, with application to an Internet-based alcohol trial *Statistics in Medicine.* 2011;30:3192-3207.

[7] Hong Jin-Liern, Jonsson Funk Michele, LoCasale Robert, et al. Generalizing Randomized Clinical Trial Results: Implementation and Challenges Related to Missing Data in the Target Population *American Journal of Epidemiology*. 2017;187:817-827.

[8] Giganti Mark J, Shepherd Bryan E. Multiple-Imputation Variance Estimation in Studies With Missing or Misclassified Inclusion Criteria *American Journal of Epidemiology.* 2020;189:1628-1632.

[9] Rubin Donald B.. Inference and Missing Data *Biometrika.* 1976;63:581-592.

[10] Snoep Jaapjan D, Morabia Alfredo, Hernández-Díaz Sonia, Hernán Miguel A, Vandenbroucke Jan P. Commentary: A structural approach to Berksons fallacy and a guide to a history of opinions about it *International Journal of Epidemiology.* 2014;43:515-521.

[11] Westreich Daniel, Daniel Rhian M.. Commentary: Berksons fallacy and missing data *International Journal of Epidemiology.* 2014;43:524-526.

[12] Choi J, Dekkers OM, Cessie S. A comparison of different methods to handle missing data in the context of propensity score analysis *Eur J Epidemiol.* 2019;34(1):23-36.

[13] van Buuren Stef, Groothuis-Oudshoorn Karin. mice: Multivariate Imputation by Chained Equations in R *Journal of Statistical Software.* 2011;45:1-67.

[14] White Ian R., Royston Patrick, Wood Angela M.. Multiple imputation using chained equations: Issues and guidance for practice *Statistics in Medicine.* 2011;30:377-399.

[15] Bareinboim Elias, Pearl Judea. Causal inference and the data-fusion problem *Proceedings of the National Academy of Sciences.* 2016;113:7345–7352.

[16] Dekkers O M, Elm E von, Algra A, Romijn J A, Vandenbroucke J P. How to assess the external validity of therapeutic trials: a conceptual approach *International Journal of Epidemiology.* 2009;39:89-94.

[17] Wang Aolin, Nianogo Roch, Arah Onyebuchi. G-computation of average treatment effects on the treated and the untreated *BMC Medical Research Methodology.* 2017;17.

[18] Tilling Kate, Williamson Elizabeth J., Spratt Michael, Sterne Jonathan A C, Carpenter James R.. Appropriate inclusion of interactions was needed to avoid bias in multiple imputation *Journal of Clinical Epidemiology.* 2016;80:107–115.

[19] Schomaker Michael, Heumann Christian. Bootstrap inference when using multiple imputation *Statistics in Medicine.* 2018;37:2252-2266.

[20] Brand Jaap P L, Buuren S., Cessie S., Hout W. B.. Combining multiple imputation and bootstrap in the analysis of cost effectiveness trial data *Statistics in Medicine.* 2019;38:210 - 220.

[21] Pattengale Nicholas D., Alipour Masoud, Bininda-Emonds Olaf R.P., Moret Bernard M.E., Stamatakis Alexandros. How Many Bootstrap Replicates Are Necessary? *Journal of Computational Biology.* 2010;17:337-354.

[22] R Core Team . *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical ComputingVienna, Austria 2019.

[23] Tompsett Daniel Mark, Leacy Finbarr, Moreno-Betancur Margarita, Heron Jon, White Ian R.. On the use of the not at random fully conditional specification (NARFCS) procedure in practice *Statistics in Medicine.* 2018;37:2338-2353.

[24] Tompsett Daniel, Sutton Stephen, Seaman Shaun R., White Ian R.. A general method for

ORIGINAL UNEDITED MANUSCRIPT

Downloaded from https://academic.oup.com/aje/advance-article/doi/10.1093/aje/kwac202/6887820 by guest on 03 January 2023

elicitation, imputation, and sensitivity analysis for incomplete repeated binary data *Statistics in Medicine.* 2020;39:2921-2935.

[25] Morris Tim P., White Ian R., Crowther Michael J.. Using simulation studies to evaluate statistical methods *Statistics in Medicine.* 2019;38:2074-2102.

[26] NHS . RSV Hospitalisaions Data https://digital.nhs.uk/data-and-information/publications/statistical/hospital-admitted-patient-care-activity/2017-18. Accessed: 12-02-2019.

[27] Zylbersztejn Ania, Almossawi Ofran, Gudka Nikesh, et al. Access to palivizumab among children at high risk of respiratory syncytial virus complications in English hospitals *British Journal of Clinical Pharmacology*. .

[28] Department of Health . *Green Book: Immunisation against Infectious Disease*. London, United Kingdom: The Stationary Office September 2015. Respiratory Synctial Virus, Chapter 27a, https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment $_{d}ata/file/458469/$ $Green_{B}ook_{C}hapter_{2}7a_{v}2_{0}W.PDF$ (visited 2021-07-1).

[29] Robins J, Hernán M A, Brumback B. Marginal structural models and causal inference in epidemiology *Epidemiology.* 2000;5:550-560.

[30] Vansteelandt S, Sjölander A. Revisiting g-estimation of the Effect of a Time-varying Exposure Subject to Time-varying Confounding *Epidemiol.Methods.* 2016;5:37-56.

[31] Zylbersztejn Ania, Gilbert R, Hjern A, Wijlaars L, Hardelid Pia. Child mortality in England compared with Sweden: a birth cohort study *The Lancet.* 2018;391:2008-2018.

[32] Yang Shu, Zhang Yilong, Liu Guanghan, Guan Qian. SMIM: A unified framework of Survival sensitivity analysis using Multiple Imputation and Martingale *Biometrics.* 2021.

[33] Taylor T, Zhou X H. Multiple Imputation Methods for Treatment Noncompliance and Nonresponse in Randomized Clinical Trials *Biometrics.* 2009;65:88–95.

[34] Corder N, Yang S. Estimating Average Treatment Effects Utilizing Fractional Imputation when Confounders are Subject to Missingness: *Journal of Causal Inference.* 2020;8:249–271.

[35] Seaman Shaun, White Ian. Review of inverse probability weighting for dealing with missing data *Statistical methods in medical research.* 2013;22:278-295.

ORIGINAL UNEDITED MANUSCRIPT

Downloaded from https://academic.oup.com/aje/advance-article/doi/10.1093/aje/kwac202/6887820 by guest on 03 January 2023

Table 1. Summary statistics of simulated variables in TT_{obs} for $n = 1,000,000$ for selected values of μ, α and γ

μ	α	γ	$P_{R_E \mid I_E=1}$ [a]	$\bar{E}$	$\bar{L}_1$	$\bar{L}_2$	$\bar{L}_3$	$\rho_{(1,2)}$ [b]	$\rho_{(2,3)}$
No Missingness									
0	0	0	1.00	1.38	0.46	0.46	0.0	-0.27	0.0
MCAR									
1.5	0	0	0.82	1.38	0.46	0.46	0.0	-0.27	0.0
0	0	0	0.50	1.38	0.46	0.46	0.0	-0.27	0.0
MAR									
1.5	0.4	0	0.83	1.40	0.45	0.51	0.07	-0.27	-0.02
1.5	0.2	0	0.83	1.39	0.45	0.49	0.04	-0.27	-0.01
1.5	-0.2	0	0.80	1.37	0.47	0.43	-0.04	-0.27	-0.01
1.5	-0.4	0	0.78	1.35	0.48	0.39	-0.08	-0.27	-0.02
0	0.4	0	0.54	1.44	0.42	0.60	0.17	-0.25	-0.03
0	0.2	0	0.52	1.41	0.44	0.53	0.09	-0.26	-0.01
0	-0.2	0	0.48	1.34	0.48	0.38	-0.10	-0.28	-0.01
0	-0.4	0	0.46	1.31	0.50	0.30	-0.20	-0.28	-0.03
MNAR									
1.5	0.4	0.4	0.88	1.43	0.46	0.51	0.05	-0.26	-0.02
1.5	0.2	0.2	0.86	1.42	0.46	0.49	0.03	-0.26	-0.01
1.5	-0.2	-0.2	0.75	1.31	0.45	0.40	-0.05	-0.29	-0.01
1.5	-0.4	-0.4	0.67	1.20	0.44	0.32	-0.12	-0.31	-0.04
1.5	0	-0.4	0.71	1.25	0.42	0.42	0.00	-0.30	0.00
1.5	0.4	-0.4	0.42	1.23	0.34	0.55	0.22	-0.29	-0.02
1.5	-0.4	0.4	0.58	1.49	0.55	0.39	-0.16	-0.24	-0.02
1.5	0.2	-0.2	0.46	1.31	0.40	0.50	0.11	-0.28	-0.01
1.5	-0.2	0.2	0.55	1.45	0.51	0.42	-0.09	-0.26	-0.00
0	0.4	-0.4	0.74	1.31	0.40	0.50	0.10	-0.28	-0.01
0	-0.4	0.4	0.85	1.42	0.49	0.43	-0.06	-0.26	0.00
0	0.2	-0.2	0.78	1.35	0.44	0.48	0.04	-0.27	0.00
0	-0.2	0.2	0.84	1.40	0.48	0.44	-0.03	-0.26	0.00
0	0.4	0.4	0.66	1.54	0.48	0.60	0.13	-0.25	-0.04
0	0.2	0.2	0.59	1.49	0.47	0.55	0.08	-0.25	-0.01
0	-0.2	-0.2	0.41	1.22	0.45	0.33	-0.11	-0.31	-0.01
0	-0.4	-0.4	0.34	1.07	0.44	0.19	-0.24	-0.34	-0.04
0	0	-0.4	0.37	1.13	0.38	0.38	0.00	-0.32	0.00

TT_{obs}:Observed Target Trial, MCAR:Missing Completely at Random, MAR:Missing at Random, MNAR:Missing not at Random.

[a] Note that $P_{R_E|I_E=1} = Pr(R_E = 1|I_E = 1)$ is a measure of the number of missing eligible.

[b] $\rho_{1,2} = \text{Corr}(L_1, L_2)$; $\rho_{2,3} = \text{Corr}(L_2, L_3)$.

ORIGINAL UNEDITED MANUSCRIPT

Downloaded from https://academic.oup.com/aje/advance-article/doi/10.1093/aje/kwac202/6887820 by guest on 03 January 2023

Table 2. Results of applying the three strategies to data generated under different scenarios, with $\boldsymbol{\mu = 1.5}$ and $\boldsymbol{n = 1,000}$.

Strategy	α[a]	γ	Bias	Coverage	RMSE	MCE
No Missingness						
1	0	0	0.00	95.0	0.00	0.01
MCAR						
1	0	0	0.01	94.4	0.17	0.01
2			0.01	93.9	0.1	0.00
3			-0.01	95.4	0.17	0.01
MAR						
1	-0.4	0	-0.04	94.8	0.20	0.01
2			-0.03	93.4	0.14	0.00
3			0.00	95.9	0.17	0.01
1	0.4	0	0.02	94.2	0.17	0.01
2			0.03	93.1	0.10	0.00
3			-0.01	95.3	0.17	0.01
MNAR						
1	0.4	0.4	0.05	93.9	0.17	0.01
2			0.06	91.6	0.14	0.00
3			-0.01	95.2	0.05	0.01
1	0.2	0.2	0.04	94.6	0.17	0.01
2			0.04	92.2	0.10	0.00
3			-0.01	95.4	0.17	0.01
1	-0.2	-0.2	-0.07	93.1	0.20	0.01
2			-0.07	89.9	0.14	0.00
3			0.01	95.7	0.05	0.01
1	-0.4	-0.4	-0.19	82.4	0.26	0.01
2			-0.17	70.4	0.22	0.00
3			0.00	96.9	0.20	0.01

MCAR:Missing Completely at Random, MAR:Missing at Random, MNAR:Missing not at Random, RMSE:Root Mean Square Error, MCE:Monte Carlo Error.

[a] Average size of TT_{obs} for the seven settings of α and γ are n= 410, 387, 415, 442, 430, 374 and 332 respectively. Average size of TT_{true} is 500. Note that $ACE^{I_E=1}$ was calculated from a single simulation with $n = 1{,}000{,}000$ and was estimated at 2.386.

ORIGINAL UNEDITED MANUSCRIPT

Downloaded from https://academic.oup.com/aje/advance-article/doi/10.1093/aje/kwac202/6887820 by guest on 03 January 2023

Table 3. Results of applying the three strategies to data generated under different scenarios, with $\mu = 0$ and $n = 1,000$.

Strategy	αa	γ	Bias	Coverage	RMSE	MCE
No Missingness						
1	0	0	0.00	95.0	0.00	0.01
MCAR						
1	0	0	0.01	95.0	0.24	0.01
2			0.01	94.6	0.14	0.00
3			0.00	97.9	0.22	0.01
MAR						
1	-0.4	0	-0.09	93.2	0.26	0.01
2			-0.07	91.7	0.17	0.00
3			-0.01	97.9	0.24	0.01
1	0.4	0	0.05	94.3	0.22	0.01
2			0.07	92.8	0.14	0.00
3			0.00	97.6	0.22	0.01
MCAR						
1	0.4	0.4	0.16	86.3	0.26	0.01
2			0.17	73.0	0.22	0.00
3			-0.00	97.3	0.20	0.01
1	0.2	0.2	0.12	90.4	0.24	0.01
2			0.12	85.3	0.17	0.00
3			0.00	97.6	0.22	0.01
1	-0.2	-0.2	-0.17	90.0	0.30	0.01
2			-0.15	83.3	0.22	0.00
3			0.00	98.2	0.07	0.01
1	-0.4	-0.4	-0.33	75.3	0.42	0.01
2			-0.31	53.4	0.34	0.01
3			0.01	98.7	0.17	0.01

MCAR:Missing Completely at Random, MAR:Missing at Random, MNAR:Missing not at Random, RMSE:Root Mean Square Error, MCE:Monte Carlo Error.

[a]Average size of TT_{obs} for the seven settings of α and γ are n= 250, 228, 271, 328, 294, 206 and 172 respectively. Average size of TT_{true} is 500. Note that $ACE^{I_E=1}$ was calculated from a single simulation with $n = 1{,}000{,}000$ and was estimated at 2.386.

ORIGINAL UNEDITED MANUSCRIPT

Downloaded from https://academic.oup.com/aje/advance-article/doi/10.1093/aje/kwac202/6887820 by guest on 03 January 2023

Table 4. Estimate of the ACE for the Palivizumab case study obtained using strategies 2 (TT_{obs}) and 3 (TT_{imp}), using an outcome model controlled for confounders and propensity score.

Sensitivity	Trial Size	$\widehat{\text{ACE}}$[a]	$\widehat{\text{ACE}}$	95% CI	Mean G-age	Mean G-age
Parameters			(%)		(Treated)	(Untreated)
TT_{obs}						
NA	1,560	-0.003	-0.3%	(-0.05,0.05)	26.5	27.2
TT_{imp}						
(0,0)	2,643	-0.002	-0.2%	(-0.04,0.04)	26.9	27.7
(-4,0)	3,659	-0.010	-1.0%	(-0.04,0.03)	26.9	25.5
(0,-4)	2,985	0.013	1.3%	(-0.03,0.05)	24.2	27.7
(-4,-4)	3,964	0.006	0.6%	(-0.02,0.04)	24.2	25.5

TT_{obs}: Target Trial emulated from observed data, TT_{imp}: Target Trial emulated from observed and imputed data, ACE: Average Causal Effect of treatment, G-age: Gestational Age.

[a]The ACE is expressed as a risk difference both in absolute value and in percentage risk difference

The sensitivity parameters are listed in order (δ_0, δ_1).

ORIGINAL UNEDITED MANUSCRIPT

Downloaded from https://academic.oup.com/aje/advance-article/doi/10.1093/aje/kwac202/6887820 by guest on 03 January 2023

Table 5. Estimated ACE for the Palivizumab case study obtained using strategies 2 (TT_{obs}) and 3 (TT_{imp}), using an Inverse Probability Weighted Outcome Model.

Sensitivity	Trial Size	$\widehat{\text{ACE}}^a$	$\widehat{\text{ACE}}$	95% CI	Mean G-age	Mean G-age
Parameters			(%)		(Treated)	(Untreated)
TT_{obs}						
NA	1,560	-0.010	-1.0%	(-0.06,0.04)	26.5	27.2
TT_{imp}						
(0,0)	2,643	-0.001	-0.1%	(-0.04,0.04)	26.9	27.7
(-4,0)	3,659	-0.031	-3.1%	(-0.08,0.01)	26.9	25.5
(0,-4)	2,985	0.023	2.3%	(-0.03,0.07)	24.2	27.7
(-4,-4)	3,964	0.011	(1.1%)	(-0.03,0.05)	24.2	25.5

TT_{obs}: Target Trial emulated from observed data, TT_{imp}: Target Trial emulated from observed and imputed data, ACE: Average Causal Effect of treatment, G-age: Gestational Age.
[a]The ACE is expressed as a risk difference both in absolute value and in percentage risk difference
The sensitivity parameters are listed in order (δ_0, δ_1).

ORIGINAL UNEDITED MANUSCRIPT

Downloaded from https://academic.oup.com/aje/advance-article/doi/10.1093/aje/kwac202/6887820 by guest on 03 January 2023

Figure 1. Directed Acyclic Graph of the assumed relationships between exposure, outcome, confounders, and data missingness indicator. A and Y are the exposure and outcome respectively. L_1 are confounders of the association between A and Y, with E the variable which determined eligibility into the target trial. L_2 and L_3 are drivers of missing data in $\boldsymbol{E}$.

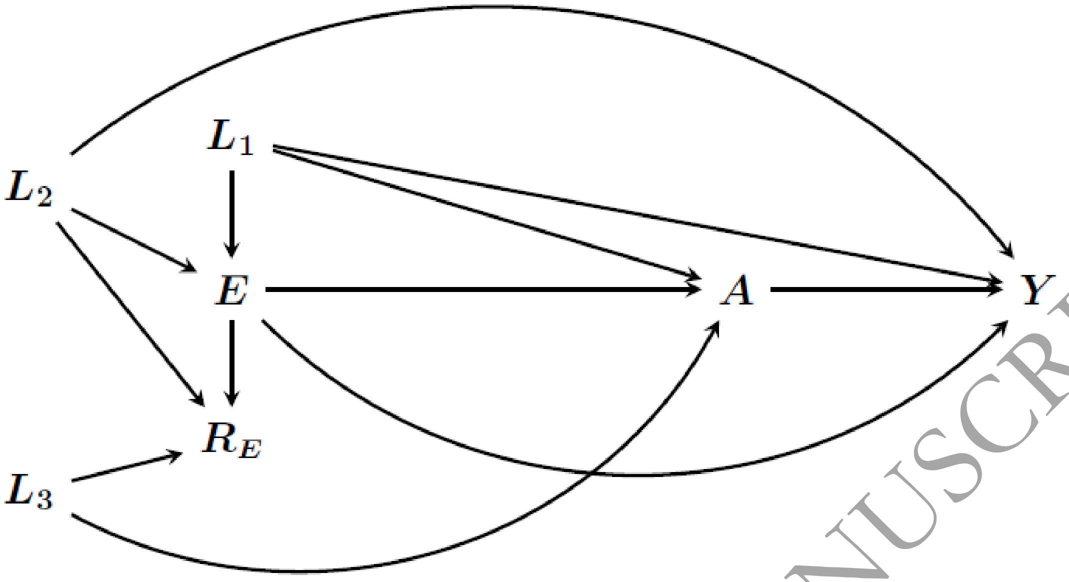
Figure 2. Directed Acyclic Graph of the assumed relationships between exposure, outcome, and confounders, and the eligibility processes represented by the indicator $I_{\boldsymbol{E}}$ plus the missing mechanism in $\boldsymbol{E}$ represented by $R_{\boldsymbol{E}}$. The solid and dashed boxes around these main indicators represent conditioning and the dotted lines represent spurious associations caused by this conditioning.

Figure 3. Derivation of the source population for the IQVIA cohort; Infants born in England between 1st Jan 2010 and 31st December 2016 with linked Hospital Episodes Statistics (HES) and prescription data. Note that the Palivizumab prescriptions database is a separate but overlapping population to those in the HTI database. Thus this population is denoted by $\boldsymbol{t}$ until linked to individuals in the HTI database population ($\boldsymbol{n}$).

Figure 4. Derivation of the complete records target trial and of the imputed target trials of Palivizumab treatment; Infants born in England between 1st Jan 2010 and 31st December 2016 with linked HES and prescription data who were eligible to receive treatment under Criteria 1a and 2a. Note the exact size of the Imputed Target Trial is unknown, and depends on the imputed data, but must be at least of size 1753 (those with complete eligible data who qualify).

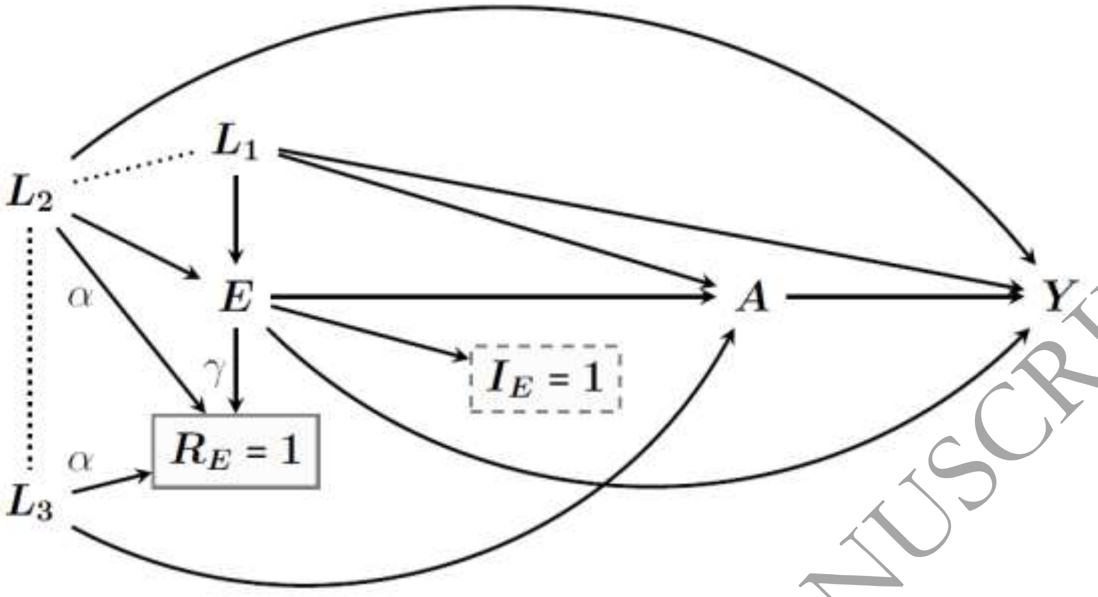
ORIGINAL UNEDITED MANUSCRIPT

Downloaded from https://academic.oup.com/aje/advance-article/doi/10.1093/aje/kwac202/6887820 by guest on 03 January 2023



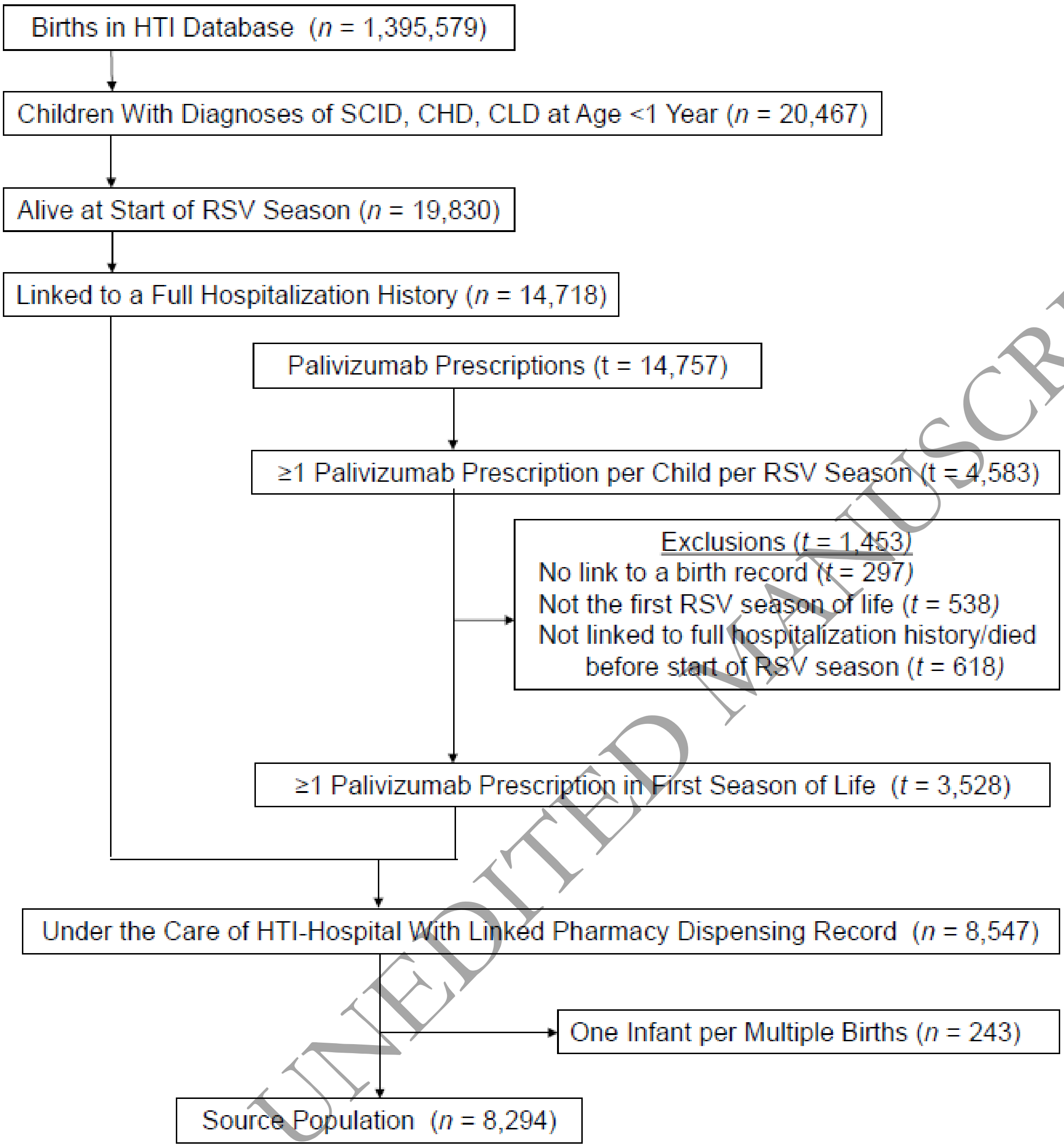
ORIGINAL UNEDITED MANUSCRIPT

Downloaded from https://academic.oup.com/aje/advance-article/doi/10.1093/aje/kwac202/6887820 by guest on 03 January 2023



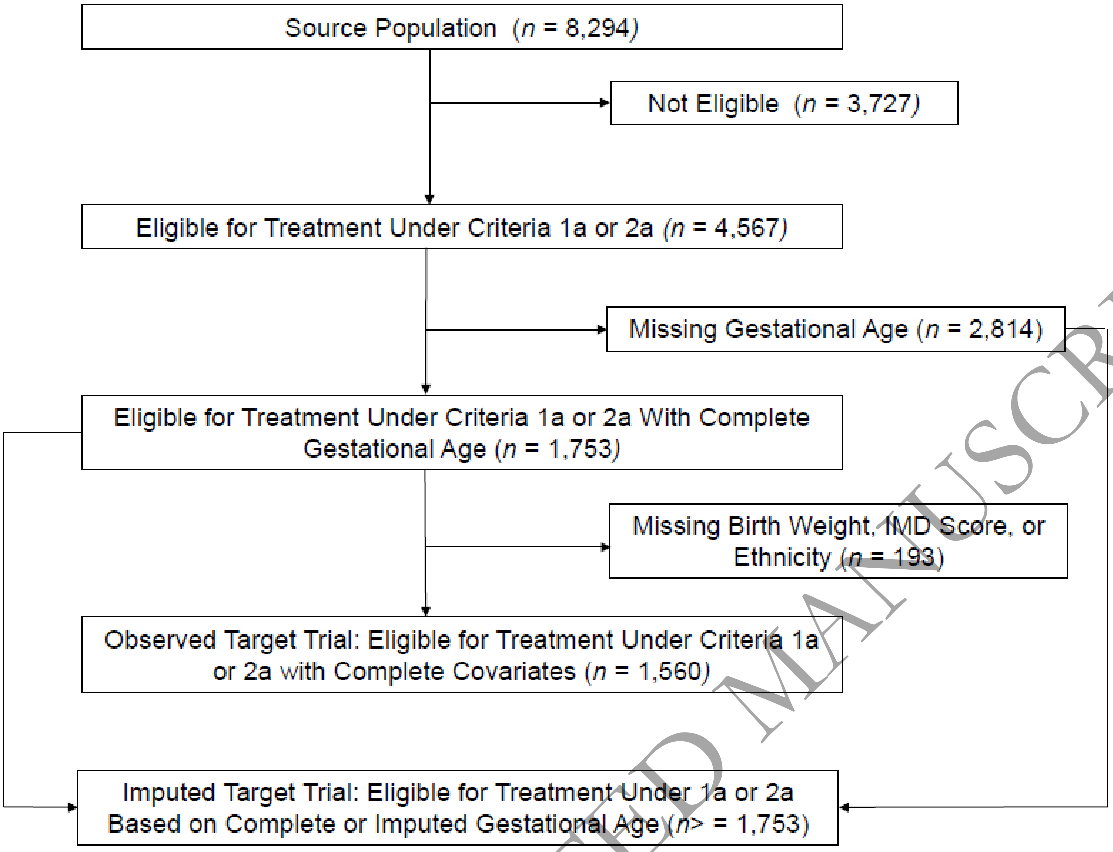
ORIGINAL UNEDITED MANUSCRIPT

Downloaded from https://academic.oup.com/aje/advance-article/doi/10.1093/aje/kwac202/6887820 by guest on 03 January 2023


Births in HTI Database (*n* = 1,395,579)
Children With Diagnoses of SCID, CHD, CLD at Age <1 Year (*n* = 20,467)
Alive at Start of RSV Season (*n* = 19,830)
Linked to a Full Hospitalization History (*n* = 14,718)
Palivizumab Prescriptions (t = 14,757)
≥1 Palivizumab Prescription per Child per RSV Season (t = 4,583)
Exclusions (*t* = 1,453)
No link to a birth record (*t* = 297)
Not the first RSV season of life (*t* = 538)
Not linked to full hospitalization history/died before start of RSV season (*t* = 618)
≥1 Palivizumab Prescription in First Season of Life (*t* = 3,528)
Under the Care of HTI-Hospital With Linked Pharmacy Dispensing Record (*n* = 8,547)
One Infant per Multiple Births (*n* = 243)
Source Population (*n* = 8,294)

ORIGINAL UNEDITED MANUSCRIPT

Downloaded from https://academic.oup.com/aje/advance-article/doi/10.1093/aje/kwac202/6887820 by guest on 03 January 2023


Source Population (*n* = 8,294)
Not Eligible (*n* = 3,727)
Eligible for Treatment Under Criteria 1a or 2a (*n* = 4,567)
Missing Gestational Age (*n* = 2,814)
Eligible for Treatment Under Criteria 1a or 2a With Complete Gestational Age (*n* = 1,753)
Missing Birth Weight, IMD Score, or Ethnicity (*n* = 193)
Observed Target Trial: Eligible for Treatment Under Criteria 1a or 2a with Complete Covariates (*n* = 1,560)
Imputed Target Trial: Eligible for Treatment Under 1a or 2a Based on Complete or Imputed Gestational Age (*n*> = 1,753)

ORIGINAL UNEDITED MANUSCRIPT

Downloaded from https://academic.oup.com/aje/advance-article/doi/10.1093/aje/kwac202/6887820 by guest on 03 January 2023