Received: Added at production | Revised: Added at production | Accepted: Added at production

DOI: xxx/xxxx

SUPPLEMENTARY MATERIAL

Propensity scores using missingness pattern information: a practical guide

Helen A. Blake* | Clémence Leyrat | Kathryn E. Mansfield | Shaun Seaman | Laurie A. Tomlinson | James Carpenter | Elizabeth J. Williamson

Correspondence
* Email: Helen.Blake@lshtm.ac.uk

In Section A, we prove that the missingness pattern approach (MPA) gives a consistent estimator of the average treatment effect under weaker versions of Mattei's assumptions,[1] referred to in the main text as the mSITA, CIT and CIO assumptions. In Section B, we explore the connection between the MPA and the missing indicator approach by comparing propensity score models for the two approaches. Section C describes the d-separation rule. Section D gives a brief overview of twin networks. In Section E, we present additional violations of the MPA's assumptions. In Section F, we provide R code for assessing the MPA's assumptions in a simple example and in our motivating example. Section G gives standardized differences for our motivating example.

A | VALIDITY OF THE MPA

In this appendix, we demonstrate that $E\left[\frac{ZY}{e^*}\right] = E[Y(1)]$ under the weaker versions of the assumptions presented in the text.

First, using the consistency assumption and rearranging, we have that:

$$E\left[\frac{ZY}{e^*}\right] = E\left[\frac{ZY(1)}{e^*}\right] = E\left[E\left[\frac{ZY(1)}{e^*}|X_{obs}, R\right]\right]$$
$$= E\left[\frac{1}{e^*}E\big[ZY(1)|X_{obs}, R\big]\right], \quad (1)$$

where $e^* = E[Z|X_{obs}, R]$.

Switching briefly to summation notation:

$$E\big[ZY(1)|X_{obs}, R\big] = \sum\sum zyP(Z|X_{obs}, R)P(Y(1)|Z, X_{obs}, R)$$
$$= \sum\sum zyP(Z|X_{obs}, R)\sum P(Y(1), X_{mis}|Z, X_{obs}, R)$$
$$= \sum\sum\sum zyP(Z|X_{obs}, R)P(Y(1)|Z, X_{mis}, X_{obs}, R)P(X_{mis}|Z, X_{obs}, R)$$

Using mSITA ($Z \perp Y(z)|X, R$ for $z = 0, 1$) and CIT ($Z \perp X_{mis}|X_{obs}, R$), we have:

$$E\big[ZY(1)|X_{obs}, R\big] = \sum\sum\sum zyP(Z|X_{obs}, R)P(Y(1)|X_{mis}, X_{obs}, R)P(X_{mis}|X_{obs}, R)$$
$$= \sum\sum zyP(Z|X_{obs}, R)\sum P(Y(1), X_{mis}|X_{obs}, R)$$
$$= \sum\sum zyP(Z|X_{obs}, R)P(Y(1)|X_{obs}, R)$$
$$= E\big[Z|X_{obs}, R\big]E\big[Y(1)|X_{obs}, R\big]$$

We can also show that $E\big[ZY(1)|X_{obs}, R\big] = E\big[Z|X_{obs}, R\big]E\big[Y(1)|X_{obs}, R\big]$ using mSITA with CIO ($Y(z) \perp X_{mis}|X_{obs}, R$ for $z = 0, 1$) in a similar manner. Thus, we can rewrite equation 1 as follows:

$$E\left[\frac{ZY}{e^*}\right] = E\left[\frac{1}{e^*}E\big[Z|X_{obs}, R\big]E\big[Y(1)|X_{obs}, R\big]\right].$$

Since $e^* = E[Z|X_{obs}, R]$:

$$E\left[\frac{ZY}{e^*}\right] = E\left[E\big[Y(1)|X_{obs}, R\big]\right] = E[Y(1)].$$

Similarly, we can show that $E[(1 - Z)Y/(1 - e^*)] = E[Y(0)]$. □

B | THE CONNECTION BETWEEN THE MISSINGNESS PATTERN APPROACH AND THE MISSING INDICATOR APPROACH

In this appendix, we consider the propensity score models for the MPA and the missing indicator approach to explore the connection between these approaches.

In a scenario with a single partially observed confounder X, the propensity score model for the MPA can be written as:

$$logit(P(Z=1)) = \begin{cases} \alpha_1 + \beta_1 X & \text{if } R=1 \\ \alpha_0 & \text{if } R=0 \end{cases}$$

with some parameters $\alpha_1, \beta_1, \alpha_0$.

Defining a new variable X^* which takes the value X if observed, and 0 otherwise, this can be rewritten as:

$$logit(P(Z=1)) = \alpha_0 + \beta_1 X^* R + (\alpha_1 - \alpha_0)R.$$

If X is binary, this is equivalent to creating a third category for X representing the missing values. If X is continuous, this sets missing values to 0 and adds an indicator variable for missing observations. This is exactly the missing indicator approach. If X were categorical, this could be extended to show that the MPA is similarly equivalent to adding a 'missing' category.

In a scenario with one partially observed confounder X, and one fully observed confounder C, the propensity score for the MPA can be written as:

$$\begin{aligned} logit(P(Z=1)) &= \begin{cases} \alpha_1 + \beta_1 X + \gamma_1 C & \text{if } R=1 \\ \alpha_0 + \gamma_0 C & \text{if } R=0 \end{cases} \\ &= \alpha_0 + \beta_1 X^* R + (\alpha_1 - \alpha_0)R + \gamma_0 C + (\gamma_1 - \gamma_0)CR \end{aligned}$$

with some parameters $\alpha_1, \beta_1, \gamma_1, \alpha_0, \gamma_0$.

In contrast, the propensity score model for the missing indicator approach is:

$$logit(P(Z=1)) = \alpha + \beta X^* R + \eta R + \gamma C$$

with parameters $\alpha, \beta, \eta, \gamma$.

This is the MPA model, constraining γ_1 to be equal to γ_0, i.e. the missing indicator model additionally assumes there are no CR interactions in the true propensity score model.

C | THE D-SEPARATION RULE

The d-separation rule, proposed within the context of directed acyclic graphs[2] and extended to SWITs,[3] determines whether a particular conditional dependency holds or not, under an assumed causal structure. Broadly speaking, association is transmitted through series of arrows — paths — in the assumed causal diagram.[4] A particular path will transmit association between the nodes at either end unless it contains a 'collider': a node which — in that path — has two incoming arrows. In Figure 1, the path $Z \leftarrow X \rightarrow Y(z)$ will transmit association between Z and $Y(z)$, but the path $Z \leftarrow U_Z \rightarrow R \leftarrow U_Y \rightarrow Y(z)$ will not because R is a collider in this path. Conditioning on a non-collider blocks associations through a specific path. Conversely, conditioning on a collider removes the blockage through that collider thereby allowing association to be transmitted. Introducing bias by conditioning on a collider is often termed collider bias.[5]

The d-separation rule states that two variables in the assumed causal diagram are conditionally independent given a set of variables V if for each path connecting the two variables: (i) the path contains two arrows which collide at a node in the path, and that node is neither in V, nor a cause of a variable in V; or (ii) the path has a non-collider which is in V.[2,4]

In Figure 1, there are two paths between Z and $Y(z)$: $Z \leftarrow X \rightarrow Y(z)$, and $Z \leftarrow U_Z \rightarrow R \leftarrow U_Y \rightarrow Y(z)$. If the conditioning set is $V = \{X\}$, then Z and $Y(z)$ are conditionally independent given V. This is because the first path contains a non-collider (X) which is in V (condition (ii)) and the second contains a collider (R) which is not in V (condition (i)). In contrast, Z and $Y(z)$ are not conditionally independent given $V = \{X, R\}$, because the second path then contains a collider (i.e. R) which is in V, and neither X nor R is a non-collider in this path.

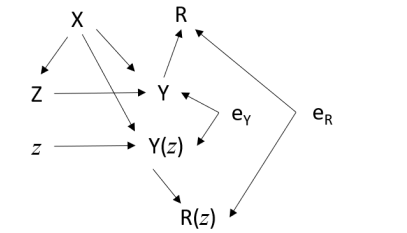

FIGURE S1 A simple example of a twin network.
X: partially observed confounder. Z: observed treatment allocation. Y: observed outcome. $Y(z)$: potential outcome resulting from intervening to set treatment to value z. R: observed missing indicator (=1 if X observed, =0 if X is missing). $R(z)$: potential missing indicator (=1 if X observed in counterfactual world, =0 if X is missing in counterfactual world). e_Y: unobserved error term between Y and $Y(z)$. e_R: unobserved error term between R and $R(z)$.

D | TWIN NETWORKS

When considering scenarios in which treatment, or the outcome, has a causal effect on missingness, by construction, the SWITs now include $R(z)$ instead of R. This means that the SWIT can no longer be used to test the MPA's assumptions. Instead, we can construct twin networks to check such scenarios, as described by Balke and Pearl,[6] and Shpitser and Pearl.[7]

Briefly, a twin network can be constructed from a directed acyclic graph, which involves real world variables and relationships, by adding counterparts of variables and relationships in the counterfactual world where treatment has been intervened upon to be set to some realisation of the random variable Z.

For example, Figure S1 shows a simple twin network of a scenario where the confounder X has a causal effect on both treatment and outcome, treatment has a causal effect on outcome, and outcome has a causal effect on missingness of the confounder. The 'real world' is shown by Z, Y, and R. The nodes z, $Y(z)$, and $R(z)$ show the counterfactual world – what would occur if we set treatment to value z. The observed outcome Y and potential outcome $Y(z)$ are connected by an unobserved error term e_Y. Similarly, the observed missing indicator R and potential missing indicator $R(z)$ are connected by an unobserved error term e_R. Because X has a causal effect on outcome, it also has a causal effect on the potential outcome. It does not, however, affect the intervened-on value of treatment, z.

To assess mSITA in Figure S1, we need to assess whether $Z \perp Y(z)|R, X$. Conditioning on X blocks the confounding pathway between Z and $Y(z)$. There is a closed path $Z \rightarrow Y \leftarrow e_Y \rightarrow Y(z)$, blocked because Y is a collider on this path. However, conditioning on R opens this path, because conditioning on a descendant of a collider (i.e. something affected by the collider) has a similar, but weaker, effect as conditioning on the collider itself. Thus the path $Z \rightarrow Y \leftarrow e_Y \rightarrow Y(z)$ is open, after conditioning on R, so the mSITA assumption may not be appropriate here. †

Dagitty can be used to assess the assumptions in twin networks, just as for SWITs.

† As d-separation for twin networks is not complete,[3] caution should be used in considering the plausibility of results that suggest two variables are not d-separated.

E | ADDITIONAL VIOLATIONS OF ASSUMPTIONS

Figure S2 summarises additional violations of the MPA's assumptions.

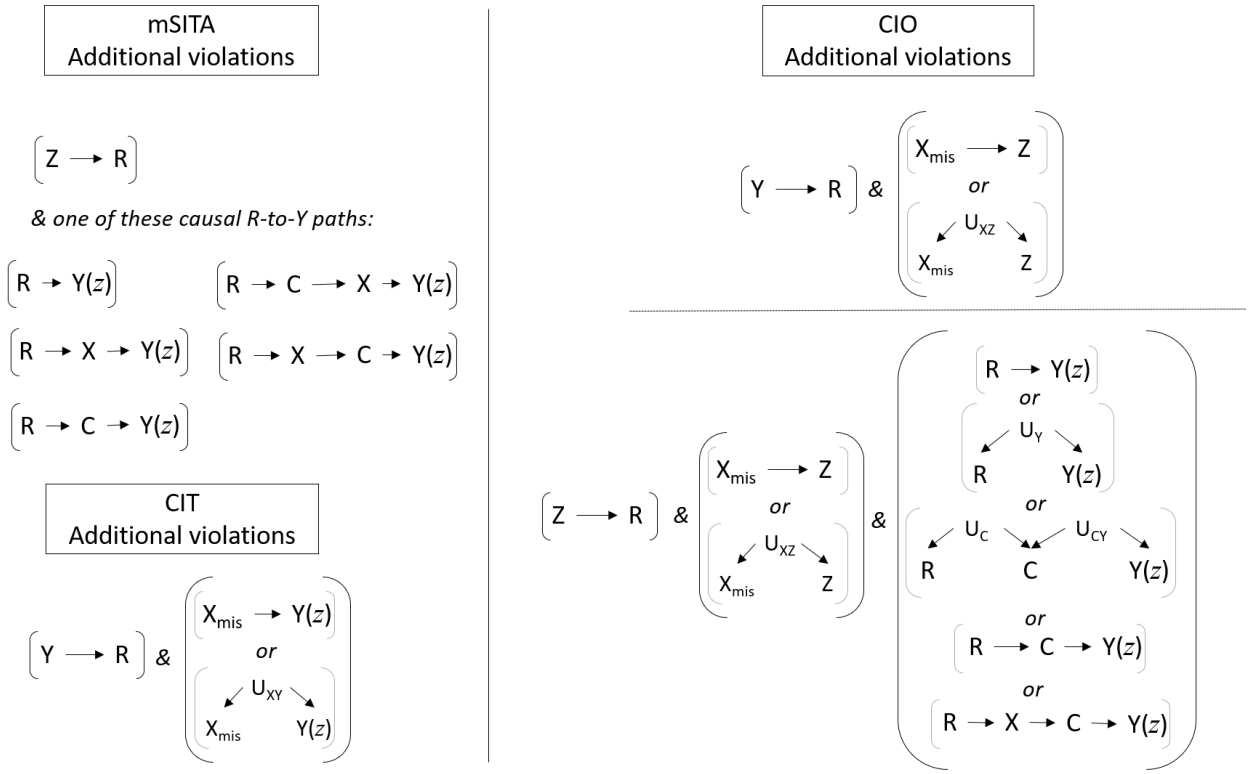

FIGURE S2 Summary of additional violations of the mSITA, CIT and CIO assumptions.
X: partially observed confounder. X_{mis}: unobserved confounder values. C: fully observed confounder. Z: treatment. $Y(z)$: potential outcome resulting from intervening to set Z equal to a particular value z. Y: observed outcome. R: missing indicator (=1 if X observed, =0 if X is missing). U_{st}: unobserved common cause between two variables s and t. U_s: unobserved common cause between R and another variable s.

F | USING DAGITTY TO ASSESS THE MPA'S ASSUMPTIONS

F.1 | Simple example: R code to use Dagitty to assess the MPA's assumptions

Run in R 3.4.0,[8] the R code below reads in our causal diagram for Figure 1 and uses d-separation to assess the mSITA, CIT and CIO assumptions.[9]

```
### R CODE TO USE DAGITTY: SIMPLE EXAMPLE


install.packages("dagitty")
library("dagitty")


############################
#   Load DAG into Dagitty  #
############################


# X       partially observed confounder
# R       observed covariate indicator: =1 if X observed, =0 otherwise.
# Z       treatment allocation (fully observed)
# Yz      potential outcome that would be observed when Z=z
# U_Y     unobserved common cause of R and Y
# U_Z     unobserved common cause of R and Z


g1 <- dagitty( 'dag {
z -> Yz
Z <- X -> Yz
R <- U_Z -> Z
R <- U_Y -> Yz
}')
coordinates( g1 ) <-
list( x=c(Z=1, z=1.2, X=2, Yz=3, R=2, U_Z=1.2, U_Y=2.8),
y=c(Z=3, z=3,   X=2, Yz=3, R=1, U_Z=1.7, U_Y=1.7) )
plot( g1 )
```

```
### Assess mSITA assumption:
### - Is Z indep of Yz given R, X, and z?
###      (note: we add z to the conditioning set because we are using a SWIG)

# List all paths between Z and Yz
paths( g1, "Z", "Yz", c("R","X","z") )

# Check whether mSITA holds
dseparated( g1, "Z", "Yz", c("R","X", "z") )

# Check whether mSITA holds if U_Y were also measured and
# included in the confounder set
dseparated( g1, "Z", "Yz", c("R","X", "z", "U_Y") )



########################################
#   DAG for subgroup with X unmeasured  #
########################################

### Suppose we believe that X does not effect Z when unmeasured
### R is now written R0 as shorthand for "R=0"

g2 <- dagitty( 'dag {
z -> Yz
X -> Yz
R0 <- U_Z -> Z
R0 <- U_Y -> Yz
}')
coordinates( g2 ) <-
list( x=c(Z=1, z=1.2, X=2, Yz=3, R0=2, U_Z=1.2, U_Y=2.8),
```

```
y=c(Z=3, z=3,   X=2, Yz=3, R0=1, U_Z=1.7, U_Y=1.7) )

plot( g2 )


### Assess CIT assumption:

### - Is Z indep of X given R=0  (and z)?

###      (note: we add z to the conditioning set because we are using a SWIG)


dseparated( g2, "Z", "X", c("R0", "z") )

paths( g2, "Z", "X", c("R0","z"))


### Assess CIO assumption:

### - Is Yz indep of X given R=0  (and z)?

###      (note: we add z to the conditioning set because we are using a SWIG)


dseparated( g2, "Yz", "X", c("R0", "z") )

paths( g2, "Yz", "X", c("R0","z"))
```

F.2 | Motivating example: R code to use Dagitty to assess the MPA's assumptions

Figure 5 shows the causal diagram which represents what the investigators believe to represent the underlying causal structure giving rise to the data. The R code below reads in our causal diagram for our motivating example and uses d-separation to assess the mSITA, CIT and CIO assumptions.

```
### R CODE TO USE DAGITTY: MOTIVATING EXAMPLE


#install.packages("dagitty")

library("dagitty")


############################

#   Load DAG into Dagitty  #

############################
```

```
# Outcome and treatment:
#   Aki   Acute Kidney Injury (outcome)
#   Ace   ACE/ARB (treatment)
#   ace   Intervened-on ACE/ARB (intervened-on treatment)

# Partially observed confounders and missing indicators:
#   Eth   Ethnicity (partially observed confounder)
#   Ckd   Baseline CKD (partially observed confounder)
#   Reth  Missingness of ethnicity
#   Rckd  Missingness of baseline CKD

# Determinants of missing data:
#   Slf   Service-level factors determining whether or not ethnicity is measured
#   Hosp  Hospitalisation

# Fully observed confounders:
#   Age
#   Sex
#   Hyp   Hypertension
#   Diab  Diabetes
#   Arr   Arrhythmia
#   Car   Cardiac failure
#   Ihd   Ischaemic heart disease

# Unmeasured factors:
#   U     (e.g. frailty)

###   Draw DAG  ###
g1 <- dagitty( 'dag {
Age -> Hyp Age -> Diab Age -> Ckd Age -> Arr Age -> Car Age -> Ihd
Sex -> Hyp Sex -> Diab Sex -> Ckd Sex -> Arr Sex -> Car Sex -> Ihd
```

```
Reth <- Eth Reth <- Slf Reth <- Hosp

Rckd <- Hyp Rckd<- Ckd Rckd <- Diab Rckd <- Age

Diab -> Ckd Ihd -> Ckd Car -> Ckd

Eth -> Arr Ihd -> Arr Arr -> Car Hyp -> Car Ihd -> Car

U -> Ckd U -> Hyp U -> Diab U -> Hosp U -> Ihd U-> Arr

Hyp -> Ace Sex -> Ace Diab -> Ace Eth -> Ace Ckd -> Ace Car -> Ace Ihd -> Ace

Age -> Aki Eth -> Aki Sex -> Aki Diab -> Aki Ckd -> Aki U -> Aki Car -> Aki

ace -> Aki

}')

coordinates( g1 ) <-

list( x=c(Age=1, Sex=1, Eth=1, Ace=6,    ace=6.5,  Aki=10,

Arr=5, Car=3.25, Ihd=5,

Reth=9,   Slf=8, Hosp=8, Hyp=3.25, Ckd=4, Diab=4, Rckd=5, U=1.5),

y=c(Age=3, Sex=5, Eth=8, Ace=4.75, ace=4.75, Aki=5,

Arr=7, Car=6.75, Ihd=6,

Reth=7.5, Slf=8, Hosp=7, Hyp=2,    Ckd=3, Diab=4, Rckd=2, U=6) )

plot( g1 )



##############################

#   Check mSITA assumption  #

##############################


### mSITA assumption:

###    Is Z indep of Yz given R, X, and z?

### Here:  Is Ace indep of Aki given Rckd, Reth, Ckd, Eth, ...

### ...Age, Sex, Hyp, Diab, Arr, Car, Ihd and ace?

###


# List all paths between Z and Yz

paths( g1, "Ace", "Aki", c("Rckd", "Reth","Ckd", "Eth",
```

```
"Age", "Sex", "Hyp", "Diab",
"Arr", "Car", "Ihd", "ace") )

# Check whether mSITA holds
dseparated( g1, "Ace", "Aki", c("Rckd", "Reth","Ckd", "Eth",
"Age", "Sex", "Hyp", "Diab",
"Arr", "Car", "Ihd", "ace") )



##########################################################
#   DAG for subgroup with CKD and ethnicity unmeasured  #
##########################################################


### Suppose we believe that:
###     Ckd does not affect prescription of ACE when unmeasured
###     Eth does not affect prescription of ACE when unmeasured

###   Draw DAG  (group with neither ethnicity nor ckd measured) ###
g2 <- dagitty( 'dag {
Age -> Hyp Age -> Diab Age -> Ckd Age -> Arr Age -> Car Age -> Ihd
Sex -> Hyp Sex -> Diab Sex -> Ckd Sex -> Arr Sex -> Car Sex -> Ihd
Reth <- Eth Reth <- Slf Reth <- Hosp
Rckd <- Hyp Rckd<- Ckd Rckd <- Diab Rckd <- Age
Diab -> Ckd Ihd -> Ckd Car -> Ckd
Eth -> Arr Ihd -> Arr Arr -> Car Hyp -> Car Ihd -> Car
U -> Ckd U -> Hyp U -> Diab U -> Hosp U -> Ihd U-> Arr
Hyp -> Ace Sex -> Ace Diab -> Ace Car -> Ace Ihd -> Ace
Age -> Aki Eth -> Aki Sex -> Aki Diab -> Aki Ckd -> Aki U -> Aki Car -> Aki
ace -> Aki
}')
```

```
coordinates( g2 ) <-
list( x=c(Age=1, Sex=1, Eth=1, Ace=6,    ace=6.5,  Aki=10,
Arr=5, Car=3.25, Ihd=5,
Reth=9,   Slf=8, Hosp=8, Hyp=3.25, Ckd=4, Diab=4, Rckd=5, U=1.5),
y=c(Age=3, Sex=5, Eth=8, Ace=4.75, ace=4.75, Aki=5,
Arr=7, Car=6.75, Ihd=6,
Reth=7.5, Slf=8, Hosp=7, Hyp=2,    Ckd=3, Diab=4, Rckd=2, U=6) )
plot( g2 )



################################
#   Check CIT/CIO assumption  #
################################



### CIT assumption:
###   Is Z indep of X given R=0  (and z)?
###   Here: is Ace indep of Ckd given Rckd=0 and Reth=0, conditional on:
###   Age, Sex, Hyp, Diab  (and ace)?
###   Here: is Ace indep of Eth given Rckd=0 and Reth=0, conditional on:
### Age, Sex, Hyp, Diab  (and ace)?

# Check whether CIT holds
dseparated( g2, "Ace", "Ckd", c("Rckd", "Reth",
"Age", "Sex", "Hyp", "Diab",
"Arr", "Car", "Ihd", "ace") )
dseparated( g2, "Ace", "Eth", c("Rckd", "Reth",
"Age", "Sex", "Hyp", "Diab",
"Arr", "Car", "Ihd", "ace") )

### CIO assumption:
```

```
###   Is Yz indep of X given R=0  (and z)?
###   Here: is Aki indep of Ckd given Rckd=0 and Reth=0, conditional on:
### Age, Sex, Hyp, Diab  (and ace)?
###   Here: is Aki indep of Eth given Rckd=0 and Reth=0, conditional on:
### Age, Sex, Hyp, Diab  (and ace)?

# Check whether CIO holds
dseparated( g2, "Aki", "Ckd", c("Rckd", "Reth",
"Age", "Sex", "Hyp", "Diab",
"Arr", "Car", "Ihd", "ace") )
dseparated( g2, "Aki", "Eth", c("Rckd", "Reth",
"Age", "Sex", "Hyp", "Diab",
"Arr", "Car", "Ihd", "ace") )


### Use similar steps to check CIT/CIO in other missingness pattern subgroups
```

G | BALANCE OF CONFOUNDERS IN MOTIVATING EXAMPLE

In Table S1, we present standardized differences [10] calculated to assess the balance of confounders in our motivating example.

TABLE S1 Standardised mean differences of confounders, before and after inverse probability of treatment weighting for complete records analysis (CRA), missingness pattern approach (MPA), missing indicator approach (MIndA), and multiple imputation (MI).
A standardized difference greater than 10% indicates imbalance for that variable.
(* Standardized differences for multiple imputation were averaged over 10 imputed datasets.)

		Percentage standardized differences (absolute values)				
Covariate		In original sample	After CRA	After MPA	After MIndA	After MI*
Age (years)	18 to 42					
	43 to 53	14.64	1.33	0.49	0.36	0.26
	54 to 62	9.42	1.64	1.51	1.45	2.26
	63 to 71	2.48	2.17	2.11	1.98	2.93
	≥ 72	4.07	2.25	3.70	3.60	4.81
Sex	Female	36.95	1.92	4.44	4.99	4.66
Chronic Kidney Disease	≤ Stage 2					
	Stage 3a	12.84	1.77	1.13	1.08	1.27
	Stage 3b	8.62	0.07	0.35	0.38	4.56
	Stage 4	3.33	0.32	0.19	0.16	1.19
Ethnicity	White					
	South Asian	6.31	0.38	0.65	0.65	7.63
	Black	3.14	3.75	3.75	4.22	8.30
	Mixed	1.73	0.56	0.69	0.84	4.25
	Other	0.43	0.42	<0.01	0.01	1.12
Diabetes Mellitus		49.21	0.43	2.70	2.01	3.06
Ischaemic Heart Disease		19.25	2.52	2.83	2.28	6.00
Arrhythmia		4.84	0.68	2.16	3.07	2.04
Cardiac Failure		32.90	1.75	0.02	0.03	0.19
Hypertension		43.08	5.74	7.85	7.88	10.93

References

1. Mattei A. Estimating and using propensity score in presence of missing background data: An application to assess the impact of childbearing on wellbeing. *Stat Methods Appl* 2009; 18(2): 257-273. doi: 10.1007/s10260-007-0086-0

2. Pearl J. Causal diagrams for empirical research. *Biometrika* 1995; 82(4): 669. doi: 10.1093/biomet/82.4.669

3. Richardson T, Robins J. Technical Report 128. Single World Intervention Graphs (SWIGs): A Unification of the Counterfactual and Graphical Approaches to Causality. http://www.csss.washington.edu/Papers/wp128.pdf; .

4. Pearl J. *Causality: Models, Reasoning and Inference*. New York: Cambridge University Press. 2nd ed. 2009.

5. Greenland S. Quantifying Biases in Causal Models: Classical Confounding vs Collider-Stratification Bias. *Epidemiology* 2003; 14(3): 300–306.

6. Balke A, Pearl J. Probabilistic Evaluation of Counterfactual Queries. In: Proceedings of the Twelfth AAAI National Conference on Artificial Intelligence; 1994: 230–237.

7. Shpitser I, Pearl J. What Counterfactuals Can Be Tested. In: Proceedings of the Twenty-Third Conference on Uncertainty in Artificial Intelligence; 2007: 352–359.

8. R Core Team . *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing; Vienna, Austria: 2017.

9. Textor J, Hardt J, Knüppel S. DAGitty: A Graphical Tool for Analyzing Causal Diagrams. *Epidemiology* 2011; 22: 745.

10. Williamson EJ, Forbes A. Introduction to propensity scores. *Respirology* 2014; 19(5): 625-635. doi: 10.1111/resp.12312