Missingness in global disaster data

EM-DAT Scientific & Technical Advisory Group (STAG) Meeting

Rebecca Jones
Email: rebecca.jones@uclouvain.be

	affected	missing	deaths	totaldeaths	reconstr_costs	insured_damages	totaldamages	ind_environment	ind_infrastrucure	ind_production
9063	2000	0	.	0	.	.	.	.	.	.
9064	17500	.	1	1	.	15000	62000	.	.	.
9065	1000000	.	25	25	.	.	1759634	.	.	.
9066	15000	.	3	3	.	.	.	.	.	.
9067	49600	.	51	51	.	.	2368508	.	.	.
9068	429	.	2	2	.	.	100000	.	.	.
9069	440000	.	37	37	.	.	49000	.	.	.
9070	225000	.	0	0	.	.	60000	.	.	.
9071	75000	.	3	3	.	.	450000	.	.	.
9072	.	.	.	0	.	.	.	.	.	.
9073	185000	.	.	0	.	.	2000	.	.	.
9074	5500000	39	37	76	.	.	1200000	.	.	.
9075	120000	.	27	27	.	.	2000	.	.	.
9076	.	.	15	15	.	.	.	.	.	.
9077	100000	.	6	6	.	.	.	.	.	.
9078	.	.	.	0	.	.	.	.	.	.
9079	9250	100	150	250	.	.	3000	.	.	.
9080	.	.	.	0	.	.	62000	.	.	.
9081	700000	.	11	11	.	.	.	.	.	.
9082	231360	.	.	0	.	.	.	.	.	.
9083	.	.	27	27	.	.	.	.	.	.
9084	.	.	.	.	.	3700000	4250000	.	.	.
9085	22545	.	70	70	.	.	15000	.	.	.
9086	5360	.	.	0	.	.	.	.	.	.
9087	.	.	6	6	.	.	.	.	.	.
9088	2000	.	15	15	.	.	377000	.	.	.
9089	10000	.	15	15	.	.	.	.	.	.
9090	350000	.	33	33	.	.	677000	.	.	.
9091	5000	.	.	0	.	.	131000	.	.	.
9092	.	.	11	11	.	.	94000	.	.	.
9093	15000000	.	30	30	.	.	925000	.	.	.
9094	150000	.	24	24	.	.	483000	.	.	.
9095	18500	.	27	27	.	.	2000	.	.	.
9096	9960000	25	56	76	.	.	6250000	.	.	.

Sources of missing disaster data

- **Unsystematic reporting** of disaster events within and across countries.
  - Differing data collection priorities.
- Technological difficulties in **disaster surveillance.**
- Methodological difficulties **quantifying disaster impacts**.
- Field-level **context**.

Consequences

3 key consequences of missing data:

1. Missing data can **bias study results**.
   - Particularly when there are systematic differences between disaster events with missing data from those with complete data.
2. Data **inefficiency**.
   - $1 - 0.91^{15} = 0.7569$ (Zhu *et al.*, 2018)
3. Reduced **external validity**.

www.nature.com/scientificdata

scientific **data**

OPEN

ANALYSIS

Human and economic impacts of natural disasters: can we trust the global data?

Rebecca Louise Jones[1,2], Debarati Guha-Sapir[3] & Sandy Tubeuf[1,2] ✉

Reliable and complete data held in disaster databases are imperative to inform effective disaster preparedness and mitigation policies. Nonetheless, disaster databases are highly prone to missingness. In this article, we conduct a missing data diagnosis of the widely-cited, global disaster database, the Emergency Events Database (EM-DAT) to identify the extent and potential determinants of missing data within EM-DAT. In addition, through a review of prominent empirical literature, we contextualise how missing data within EM-DAT has been handled previously. A large proportion of missing data was identified for disasters attributed to natural hazards occurring between 1990 and 2020, particularly on

UCLouvain

Data

- Emergency Events Database (EM-DAT).
- All disaster events attributed to natural hazards occurring between 1990 – 2020 (n = 11,124).
- Variables of interest:
  - **Total estimated damages (US$)**
  - Reconstruction costs (US$)
  - Insured Damages (US$)

Economic Losses

  - **No. of [people] Affected**
  - **No. of [people] Missing**
  - **No. of Deaths**
  - Total Deaths (No. of Deaths + No. of Affected)

Human Losses

UCLouvain

Methods

Steps involved in a missing data diagnosis:

1. Describing **proportions** of missing data.
   - STATA code: 'm d e s c'

2. Visualising missing data **patterns**.
   - STATA code: 'm i s s p a t t e r n'

3. Informing the **mechanisms** of missing data.
   - By logistic regression analysis, Little's MCAR test or univariate correlation analysis.
   - Underpins the choice of missing data method.

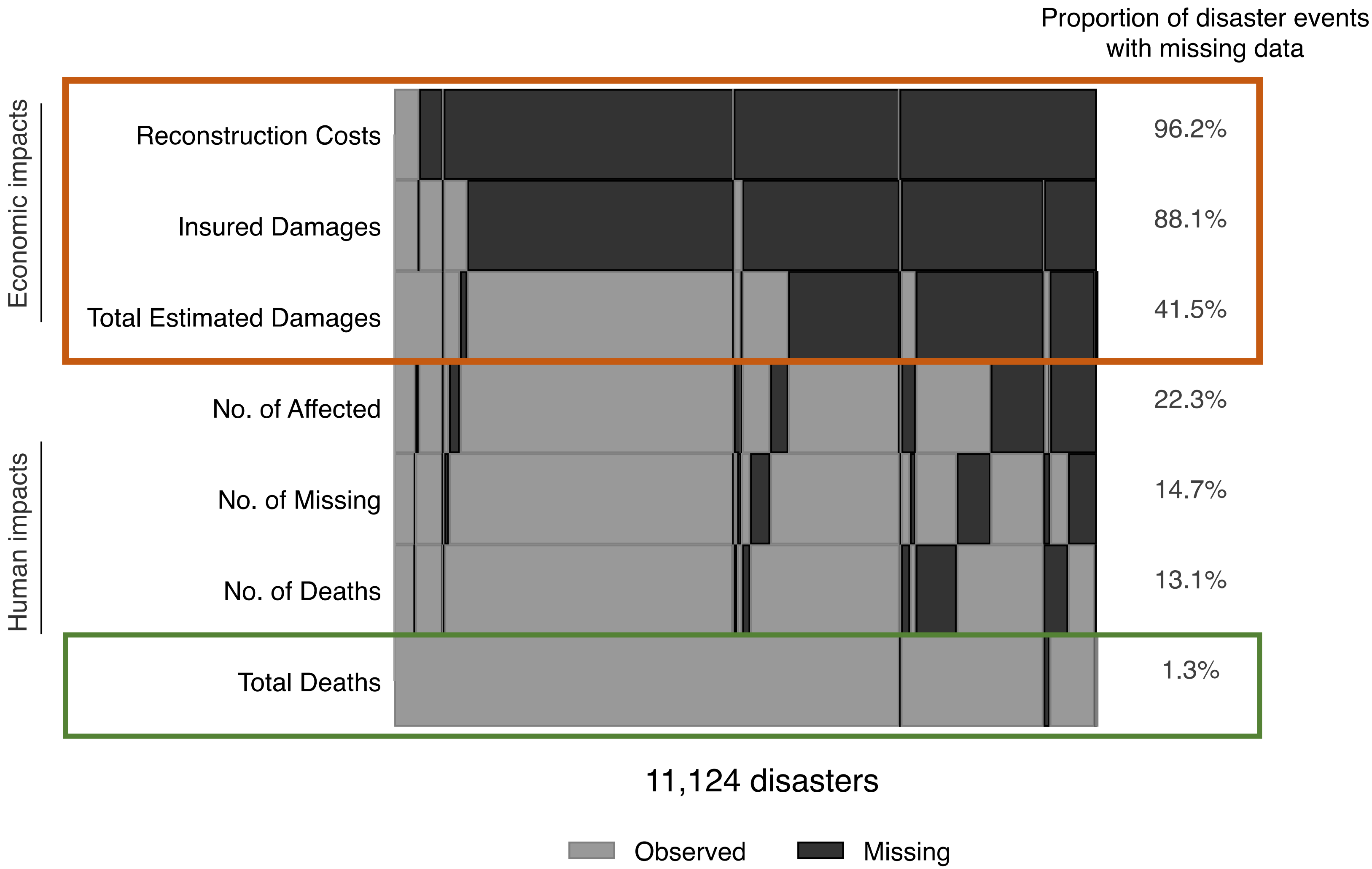
Mechanisms of missing data

Mechanisms of missing data as defined by Rubin (1976).

Missing data mechanism	**Definition**
Missing Completely At Random (MCAR)	The probability of missingness is <u>independent of</u> both <u>observed</u> and <u>unobserved</u> data.
Missing At Random (MAR)	Given the observed data, the probability of missingness <u>is independent of unobserved</u> data.
Missing Not At Random (MNAR)	The probability of missingness is <u>dependent</u> of both <u>observed</u> and <u>unobserved</u> data.

- We can test deviations from the assumption of MCAR, but not MAR.

Results



Proportion of disaster events with missing data
Economic impacts
Human impacts
Reconstruction Costs
Insured Damages
Total Estimated Damages
No. of Affected
No. of Missing
No. of Deaths
Total Deaths
96.2%
88.1%
41.5%
22.3%
14.7%
13.1%
1.3%
11,124 disasters
Observed
Missing

Results

- The observed data partially explained the probability of **Total Estimated Damages** to be missing (pseudo-R^2 = 0.416).
- Explained less the probability of **No. Affected** and **No. of Deaths** to be missing (pseudo-R^2 = 0.206, pseudo-R^2 = 0.188).

More specifically, the probability of missingness on:

- **Total Estimated Damages:**
  - ⬆ Disaster events occurring after the year 2002.
  - ⬆ Disaster events occurring in lower-income countries.
  - ⬇ Droughts, Epidemics and Extreme Temperature Events.*
  - ⬇ Higher severity disaster events.
- **No. Affected** and **No. of Deaths**:
  - ⬆ Disaster events occurring after the year 2002.
  - ⬇ For disaster events occurring in lower-income countries.

* In reference to floods

Key takeaways

- Missing data in EM-DAT is **unlikely to be MCAR**.
- Systematic change in the reporting of disaster impacts after the year 2002.
- Predictors of missingness differed for economic and human losses.
- Disaster aid might incentivise the reporting of human losses by lower-income countries.

Can we learn from the existing disaster literature?

NIHR | National Institute for Health Research

PROSPERO
International prospective register of systematic reviews

Untold story of missing data in disaster research: a systematic review of the empirical literature utilising the Emergency Events Database (EM-DAT).

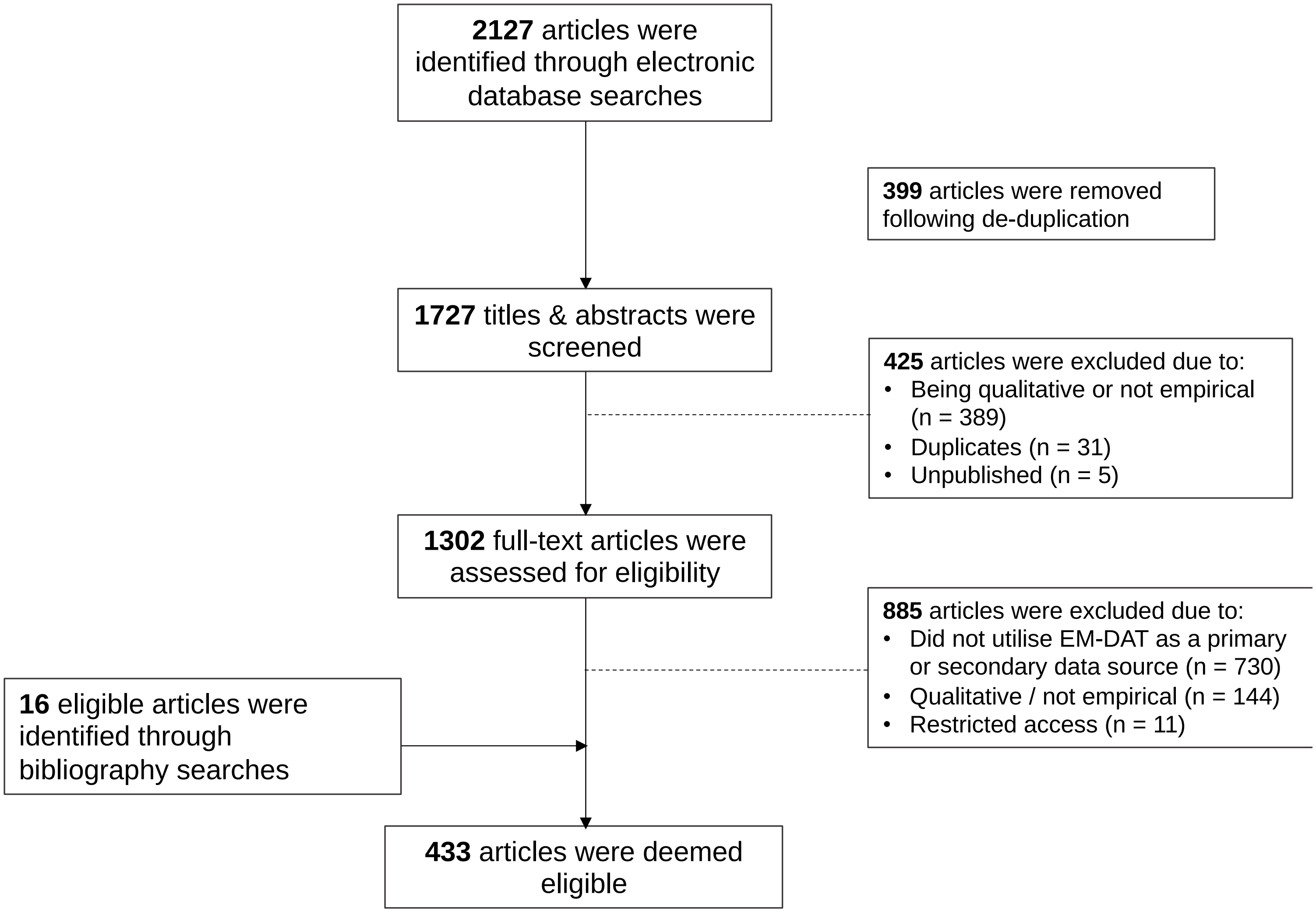
Rebecca Louise Jones[1,2], Aditi Kharb [3], Sandy Tubeuf [1,2]

- Comprehensive systematic literature review.
- Electronic database searches of:
  - EconPapers (RePEc), EconLit (Ovid), EMBASE, MEDLINE (PubMed), Web of Science, Global Health Database (EBSCOhost), The Cochrane Library, Scopus, JSTOR and Google Scholar.
- Primary research question:

  How are missing data acknowledged and handled in the empirical, quantitative literature utilising EM-DAT as a primary or secondary data source?

UCLouvain

Results



UCLouvain

Results

Acknowledging missing data:

- Of the 433 eligible studies, 200 (46.2%) studies acknowledged missing data.
    - 125 studies (62.5%) acknowledged missing data only briefly.
    - 23 studies (11.5%) attempted to diagnose the potential mechanisms of missing data.

Handling missing data:

- Of the 433 eligible studies, 145 (36.5%) attempted to handle missing data.
- 24 different approaches to handle missing data.

Results

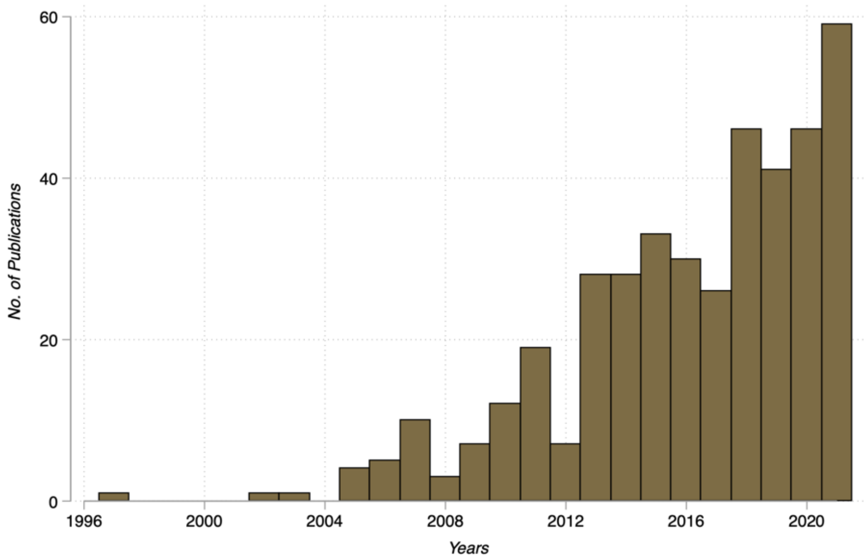
Method	Classification	Description	Frequency
Excluding observations *ad-hoc*	Deletion	Excluding select observations, or groups of observations in an *ad-hoc* manner.	30
Complete Case Analysis (CCA) (Listwise deletion)	Deletion	Excluding observations with missing data on at least one variable of interest. Also referred to as row deletion.	27
Supplementing with other data sources	Imputation	Filling data gaps with data from alternative sources either manually, or by merging data sources.	27
Restricting the scope of analysis	Deletion	Restricting the geographical or temporal scope of the analysis based on data availability.	23
Imputation (unspecified)	Imputation	Imputing missing data to generate a complete dataset.	15
Aggregating observations	Deletion	Compiling and expressing individual-level data into summary forms for statistical analysis.	11
Column Deletion	Deletion	Deleting variables which have a high proportion of missing data. A threshold of greater than 60% missing data is commonly suggested.	8
Interpolation	Augmentation	Estimating missing data values based on a known range of discrete, observed data points.	8
Zero-value Imputation	Imputation	Treating all missing data as true zero values and substituting accordingly. A type of single imputation.	7
Available Case Analysis (ACA) (Pairwise deletion)	Deletion	Utilising all observed data points for each variable, or pair of variables, to calculate sample ‘moments’ (population mean, variance, etc.). Sample moments are then included in data analysis in place of population parameters.	5

- 3 broad approaches to handle missing data.
- The most common approaches employed were *ad-hoc* with little statistical basis.

UCLouvain

So what?

- Increasing demand for global disaster data.



Use of EM-DAT in the empirical literature over the last 25 years (1996 – 2021). CRED (2022).

- Deletion methods, which assume missing data are MCAR, are frequently used in the disaster literature.
- **Raises doubt** regarding the accuracy of study results.

Potential next steps…

- Conduct a **simulation analysis** to determine:
  - What extent do missing data bias study results?
  - Which methods are most appropriate to handle missing data in disaster databases?
- Construct a suitable **framework** to guide researchers in the appropriate consideration of missing disaster data.

Supplementary Table 4. Glossary of conventional and advanced missing data methods.

Method	Description	Notes
Conventional methods		
Column deletion	Deleting variables which have a high proportion of missing data. A threshold of greater than 60% missing data is commonly suggested.	This method should only be considered for variables which are not necessary to the analysis.
Complete Case Analysis (CCA) (Listwise deletion)	Also referred to as row deletion. Observations with missing data on at least one variable of interest are excluded.	CCA is used by default in most statistical software programmes. It yields a complete dataset which facilitates the use of conventional data analysis methods. When a dataset contains a large proportion of missing data, CCA excludes a large fraction of the original data and reduces the statistical power of analyses. CCA relies on the assumption that missing data are MCAR or MAR if all predictors of missingness are included in the analysis.
Aggregating data	Compiling and expressing individual-level data into summary forms for statistical analysis.	Missing data are masked within summary statistics, minimising their relative impact. However, the precision of analyses are substantially reduced.
Dummy variable adjustment	For continuous variables, a dummy variable is created to indicate if data is missing on that variable. For categorical variables, an additional category is created to hold cases with missing data.	This method allows the entire dataset to be used in data analysis, maximising the sample size and statistical power. However, dummy variable adjustment has been shown to yield biased parameter estimates.
Available Case Analysis (ACA) (Pairwise deletion)	All observed values for each variable or pair of variables are utilised to calculate sample 'moments' (population mean, variance etc.). In other words, only missing data for the variable, or pairs of variables of interest are excluded. Sample moments are then included in the data analysis in place of population parameters.	Like CCA, this method yields a complete dataset which facilitates the use of conventional data analysis methods. As ACA uses all the data available for each analysis, it does not skew summary statistics. For bivariate and multivariate analyses, ACA requires sufficient correlation between variables to yield consistent parameter estimates. However, as different subsets of the data are used to calculate sample moments, there is no guarantee of this. ACA relies on the assumption of MCAR.
Mean imputation	Missing values are substituted with a single unconditional mean of the observed values.	Single imputation methods yield a complete dataset and facilitate the use of conventional data analysis methods independently of missing data methods. As with most single imputation methods, mean imputation yields biased parameter estimates. Predicted values do not contain random error, so sample variation is reduced. This can lead to an underestimation of standard errors and optimistic significance values. This issue is magnified with higher proportions of missing data.
Regression-based imputation	Missing values are substituted with a single, predicted value estimated using regression methods, conditional on observed predictors of missingness.	Single imputation methods yield a complete dataset and facilitate the use of conventional data analysis methods independently of missing data methods. Relies on the assumption that missing data are MAR. As with mean imputation, regression-based imputation yields biased parameter estimates and uncertainty in the predicted value is not adequately reflected. Predicted values do not contain random error, so sample variation is reduced. This can lead to an underestimation of standard errors and optimistic significance values. This issue is magnified with higher proportions of missing data.
Data merging	Merging data sources, or data subsets by integration or aggregation to supplement existing data.	Data merging by conditional merging is most appropriate when merging incomplete datasets. This involves filling missing data gaps with observed values found in other source datasets. Data loss and file-matching errors can occur if there is heterogeneity in the coding of data across datasets, or if there is heterogeneity in the number and type of variables. Hence, datasets need to be standardised prior to merging. Data matching is also necessary to prevent the duplication of data across datasets. This method can therefore be time-consuming.
Advanced methods		
Inverse probability weighting (IPW)	'Complete cases' are weighted by the inverse probability of being observed. Weights are calculated using a binary regression model conditional on observed predictors of missingness.	IPW rebalances the data so complete cases better represent the entire sample. By adjusting for missing data without manipulating the full dataset, IPW does not create issues of incompatibility with subsequent data analysis. Relies on the assumption that missing data are MCAR or MAR, if all predictors of missingness are included in the binary regression model.
Maximum-likelihood	Uses all observed data to generate the parameter estimates most likely to result from the available data. Likelihoods are computed separately for observations with complete and incomplete data on the variables of interest. The product of the individual likelihoods is then maximised to give the maximum-likelihood parameter estimates.	Maximum likelihood yields asymptotically unbiased and efficient parameter estimates. Missing data and parameter estimation are handled in a single step. However, this requires all predictors of missingness to be specified in the intended analysis model. Relies on the assumption that missing data are MAR but can be modified for missing data which are MNAR. For each variable with missing data, parametric models for the joint distributions need to be specified. This is potentially difficult and parameter estimates may be sensitive to the choice of model. Maximum-likelihood is limited to only linear models and requires specialist statistical software packages.
Multiple imputation	An extension of regression-based single imputation. Multiple imputation involves 3 steps: 1. Imputation using regression methods is performed several times, generating *m* imputed datasets. Each dataset contains a different, randomly drawn, imputed value for all missing values. 2. Datasets are analysed separately using standard methods. 3. The parameter estimates and standard errors obtained from each are combined using Rubin's Rules to generate a single set of parameter estimates and standard errors.	Multiple imputation yields asymptotically unbiased and efficient parameter estimates. By generating multiple, randomly drawn imputed values, multiple imputation adequately accounts for uncertainty in the predicted value. Makes no assumptions about the missing data mechanism; can be modified for missing data which is MNAR. Requires several decisions to be made on: the type of imputation model, the number of imputations (m), the number of iterations between imputations and the choice of prior distribution. With larger proportions of missing data, a greater number of imputations are required. Generally, $m = 20$ is considered sufficient. Potentially computationally difficult with a large number of variables and/or observations.
Other advanced methods		
Hot deck imputation	Each missing value is replaced with a plausible, observed value taken from similar observations within the same classification. Imputed values may be selected at random, or by using distance metrics, such as nearest neighbour matching.	This method yields a complete dataset and facilitates the use of conventional data analysis methods independently of missing data methods. Hot deck imputation does not require missing values to be modelled. Therefore, parameter estimates are less sensitive to model misspecifications. If there are large proportions of missing data, only a small sample of observations may be used to impute missing values, leading to replication of values and reduced sample variation. This can lead to an underestimation of standard errors and optimistic significance values.
Bayesian simulation	An extension of multiple imputation. Missing data are treated as additional, unknown variables for which posterior predictive distributions can be calculated by specifying a missing data model and Bayesian priors. Algorithms, such as Monte Carlo Markov Chain are then used to yield parameter estimates from the posterior predictive distributions.	Missing data and parameter estimation are handled in a single step. Bayesian analysis can be easily adapted for incomplete data. Bayesian priors may be based on expert opinion which can improve the reliability of results. As with multiple imputation, Bayesian simulation adequately accounts for uncertainty due to the missing values. Can be modified to account for any assumption on the mechanism of missing data Parameter estimates may be sensitive to model misspecifications. Requires specialist software and can be highly complex.

ACA, available case analysis; CCA, complete case analysis; MAR, missing at random; MCAR, missing completely at random; MNAR, missing not at random.

UCLouvain