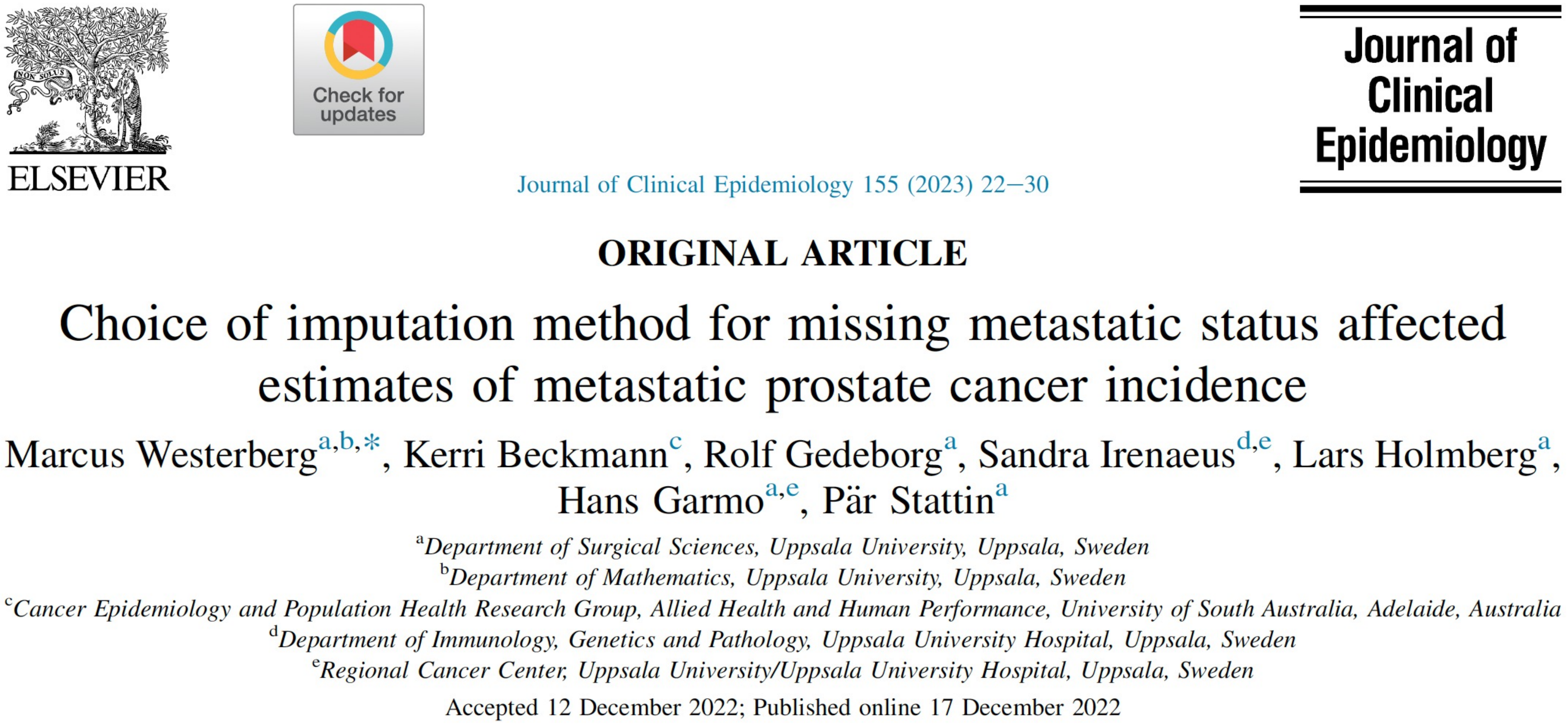


ELSEVIER

Check for updates

Journal of Clinical Epidemiology 155 (2023) 22–30

Journal of Clinical Epidemiology

ORIGINAL ARTICLE

Choice of imputation method for missing metastatic status affected estimates of metastatic prostate cancer incidence

Marcus Westerberg[a,b,*], Kerri Beckmann[c], Rolf Gedeborg[a], Sandra Irenaeus[d,e], Lars Holmberg[a], Hans Garmo[a,e], Pär Stattin[a]

[a]*Department of Surgical Sciences, Uppsala University, Uppsala, Sweden*
[b]*Department of Mathematics, Uppsala University, Uppsala, Sweden*
[c]*Cancer Epidemiology and Population Health Research Group, Allied Health and Human Performance, University of South Australia, Adelaide, Australia*
[d]*Department of Immunology, Genetics and Pathology, Uppsala University Hospital, Uppsala, Sweden*
[e]*Regional Cancer Center, Uppsala University/Uppsala University Hospital, Uppsala, Sweden*

Accepted 12 December 2022; Published online 17 December 2022

Commented by Sureerat Suwatcharangkoon

October 6, 2023

Types of missing data

Missing completely at random (MCAR)

- *The probability of being missing is the same for **all cases***

Missing at random (MAR) → Multiple imputation

- *The missingness depends on information we have already **observed***

Missing not at random (MNAR) → Sensitivity analyses

- *The probability that data are missing depends on the **unobserved data***

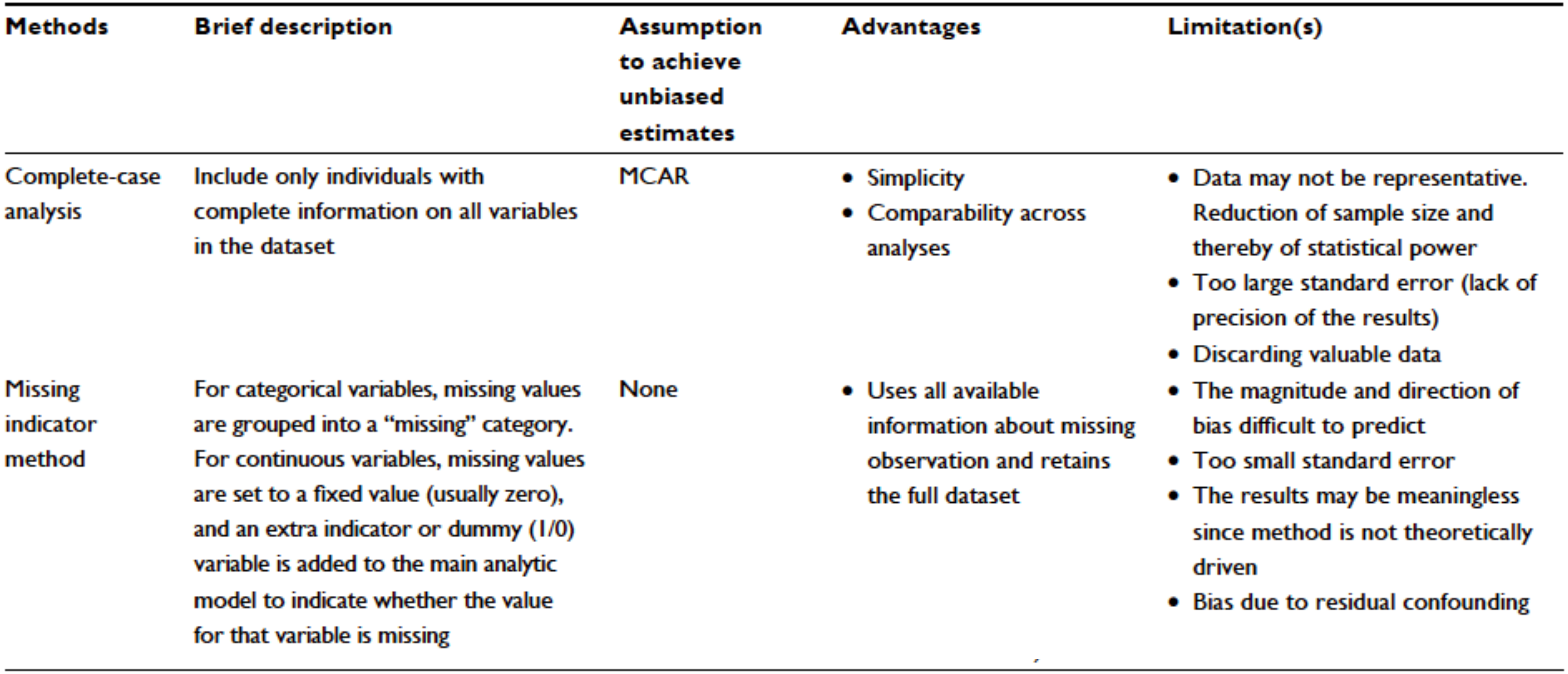
BMJ 2009;338:b2393
Clinical Epidemiology 2017:9 157–166

Proposed methods for dealing with missing data in the design phase

- Optimizing data collection
- Pilot studies can help to identify variables particularly susceptible to missing values, and steps
- Regular monitoring of data quality and completeness
- Patients may be asked to provide reasons for refusing to participate

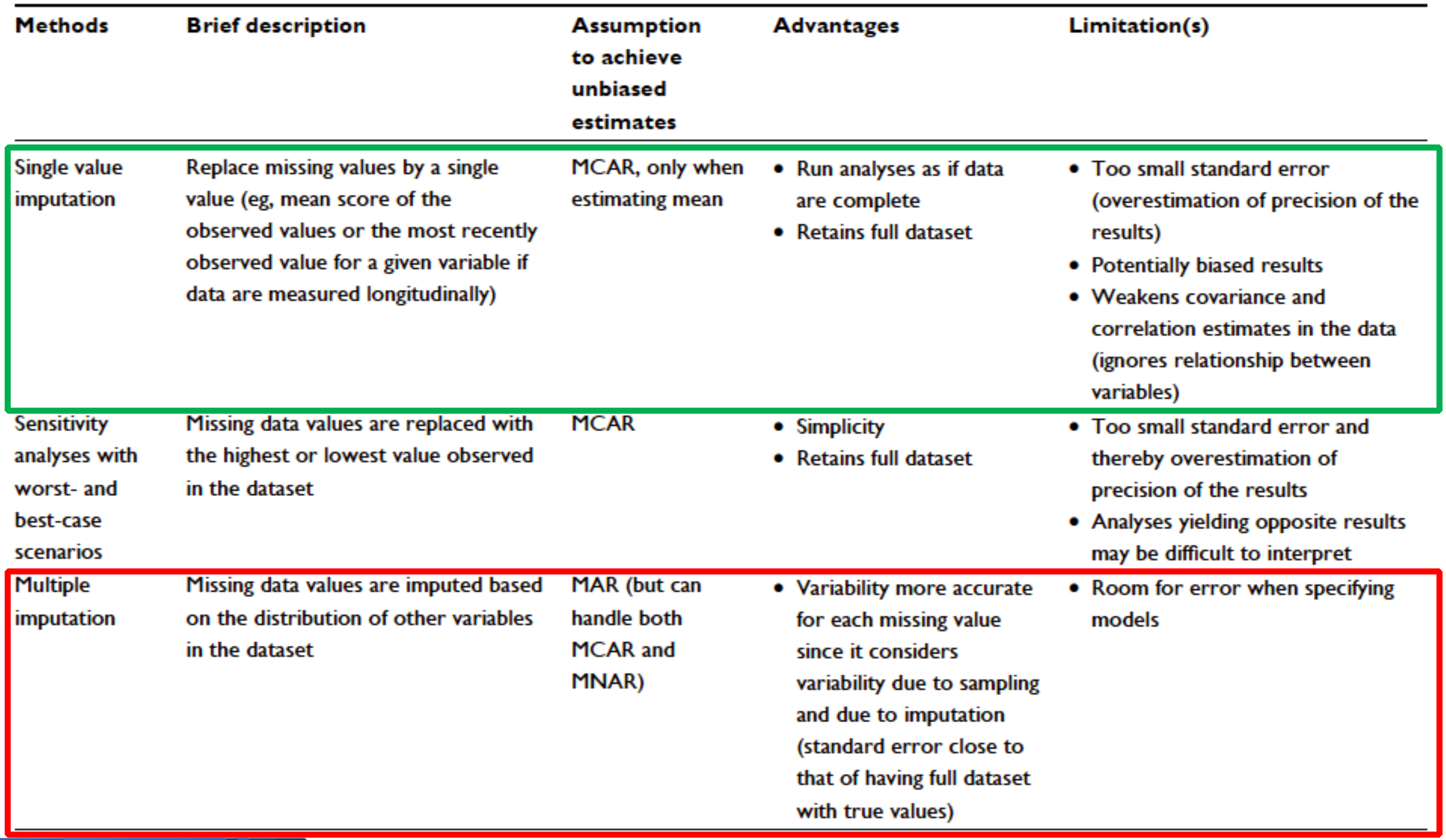
Clinical Epidemiology 2017:9 157–166

Proposed methods for dealing with missing data in the analytic phase



Methods	Brief description	Assumption to achieve unbiased estimates	Advantages	Limitation(s)
Complete-case analysis	Include only individuals with complete information on all variables in the dataset	MCAR	• Simplicity • Comparability across analyses	• Data may not be representative. Reduction of sample size and thereby of statistical power • Too large standard error (lack of precision of the results) • Discarding valuable data
Missing indicator method	For categorical variables, missing values are grouped into a "missing" category. For continuous variables, missing values are set to a fixed value (usually zero), and an extra indicator or dummy (1/0) variable is added to the main analytic model to indicate whether the value for that variable is missing	None	• Uses all available information about missing observation and retains the full dataset	• The magnitude and direction of bias difficult to predict • Too small standard error • The results may be meaningless since method is not theoretically driven • Bias due to residual confounding

Clinical Epidemiology 2017:9 157–166



Methods	Brief description	Assumption to achieve unbiased estimates	Advantages	Limitation(s)
Single value imputation	Replace missing values by a single value (eg, mean score of the observed values or the most recently observed value for a given variable if data are measured longitudinally)	MCAR, only when estimating mean	• Run analyses as if data are complete • Retains full dataset	• Too small standard error (overestimation of precision of the results) • Potentially biased results • Weakens covariance and correlation estimates in the data (ignores relationship between variables)
Sensitivity analyses with worst- and best-case scenarios	Missing data values are replaced with the highest or lowest value observed in the dataset	MCAR	• Simplicity • Retains full dataset	• Too small standard error and thereby overestimation of precision of the results • Analyses yielding opposite results may be difficult to interpret
Multiple imputation	Missing data values are imputed based on the distribution of other variables in the dataset	MAR (but can handle both MCAR and MNAR)	• Variability more accurate for each missing value since it considers variability due to sampling and due to imputation (standard error close to that of having full dataset with true values)	• Room for error when specifying models

Clinical Epidemiology 2017:9 157–166

Specification of imputation models

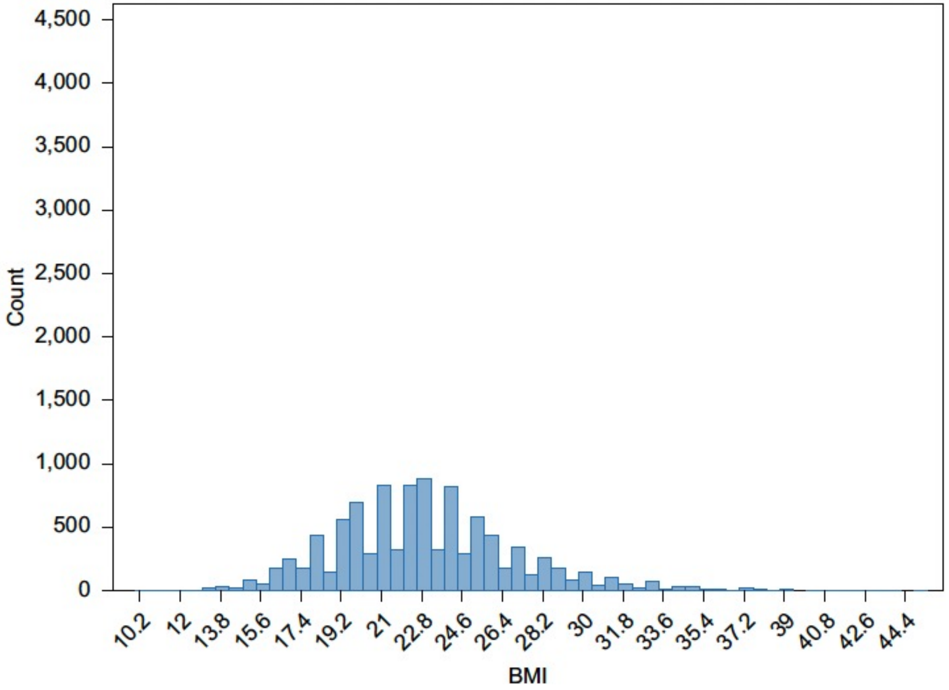
1. Deterministic imputation
2. Partial deterministic imputation + MI (PDI + MI)
3. Standard MI (SMI)
4. Restricted MI (RMI)

1. Deterministic imputation

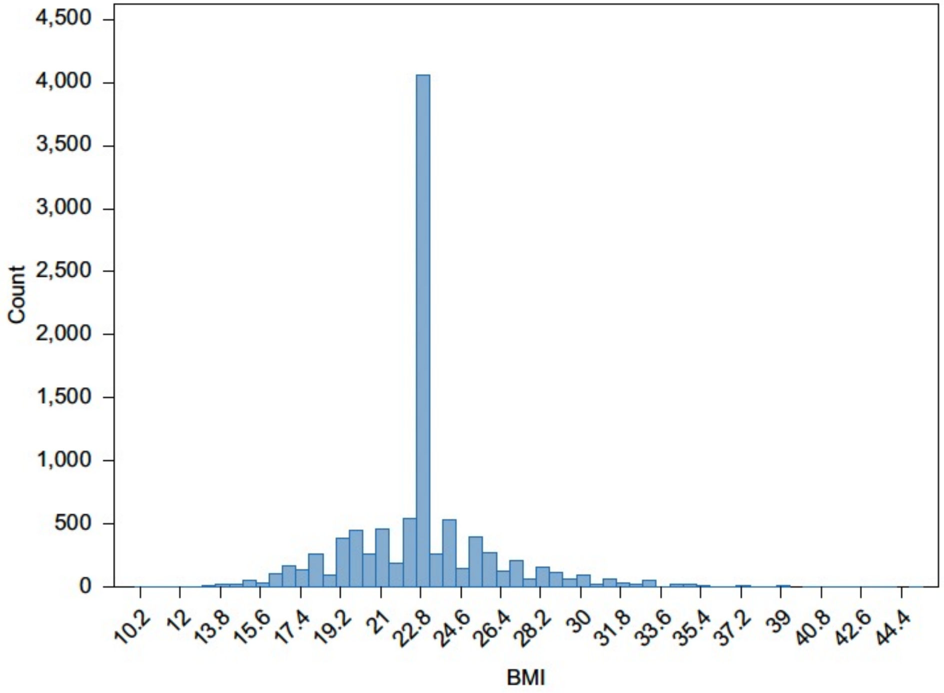
- Substitution method (missing M stage -> M0)
- Imaged men with M0 cannot be differentiated from nonimaged men
- Unbiased estimates for the population means or totals if
  - the missing values are missing completely at random (MCAR)
  - the missing values only depend on the auxiliary variables which are used to construct the imputation cells

A Study of Imputation Algorithms
Ralph Lee. U.S. Department of Education. National Center for Education Statistics.

Single value imputation



Normal distribution of observed BMI in a full dataset of 10,000 observations.



Distribution of BMI in a dataset of 10,000 observations, where 35% of BMI values are missing and replaced by the observed mean BMI value

Clinical Epidemiology 2017:9 157–166

1. Deterministic imputation

- However, the distribution of the data will be distorted substantially and the concentration of all imputed values at the cell means creates spikes in the distribution.
- Therefore, quartile estimates will be biased, and the variances materially underestimated.

A Study of Imputation Algorithms
Ralph Lee. U.S. Department of Education. National Center for Education Statistics.

1. Deterministic imputation

- Variance-covariance estimates calculation by the adjusted mean imputation (or substitution) method
- Using a denominator of n-m-1 instead of n-1

  (n = sample size, m = number of cases missing)

A Study of Imputation Algorithms
Ralph Lee. U.S. Department of Education. National Center for Education Statistics.

1. Deterministic imputation

- Cohen (1996) suggested another way to adjust variance estimates by imputing more diversified values for the missing cases.
- Imputing half of the missing values with

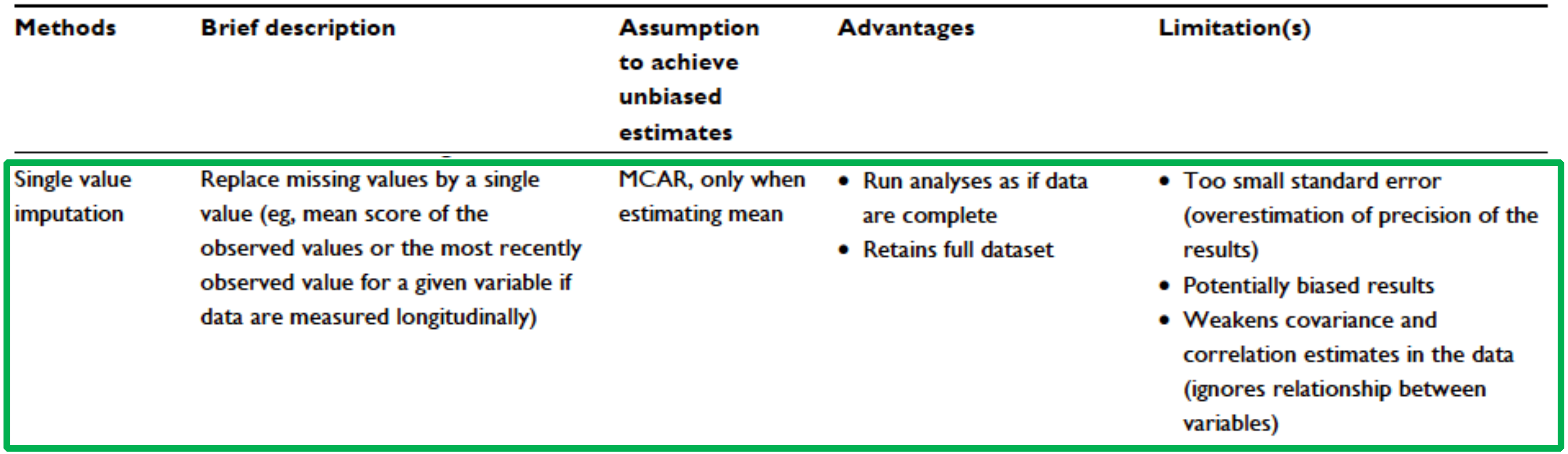
$$\bar{y}_r + \sqrt{\frac{n+r-1}{r-1}}D_r$$

$$\bar{y}_r - \sqrt{\frac{n+r-1}{r-1}}D_r$$

r = number of response values, y_r = mean of observed values,

$$D_r^2 = \frac{1}{r}\sum_1^r (y_i - \bar{y}_r)^2$$

A Study of Imputation Algorithms
Ralph Lee. U.S. Department of Education. National Center for Education Statistics.



Methods	Brief description	Assumption to achieve unbiased estimates	Advantages	Limitation(s)
Single value imputation	Replace missing values by a single value (eg, mean score of the observed values or the most recently observed value for a given variable if data are measured longitudinally)	MCAR, only when estimating mean	• Run analyses as if data are complete • Retains full dataset	• Too small standard error (overestimation of precision of the results) • Potentially biased results • Weakens covariance and correlation estimates in the data (ignores relationship between variables)

Clinical Epidemiology 2017:9 157–166

2. Partial deterministic imputation + MI (PDI + MI)

- **PDI**: Low metastatic risk with missing M stage -> M0
- **MI**: Remaining missing data in M stage and all other variables (e.g., PSA and N stage) was imputed using MI including all variables.

Journal of Clinical Epidemiology 155 (2023) 22-30

2. Partial deterministic imputation + MI (PDI + MI)

The following variables were included: age at diagnosis, year of diagnosis, and M stage, and the auxiliary variables

- logPSA, T and N stage, Gleason sum, WHO grade, primary treatment, mode of detection, follow-up time and cause of death (prostate cancer or other causes) or censoring.

Imputation based on Clinical Imputation

- Imputation missing N- and M-stage data only in low/intermediate-risk men

For example:

- T1 and Gleason 6 -> unlikely to have nodal involvement or distant metastases **(N0 and M0)**
- Staging data is available for nodal disease but missing data for distant metastases -> likely that staging was performed and very low likelihood of missing M-stage representing positive disease **(M0)**

Cancer Epidemiol. 2019 Feb:58:44-51.

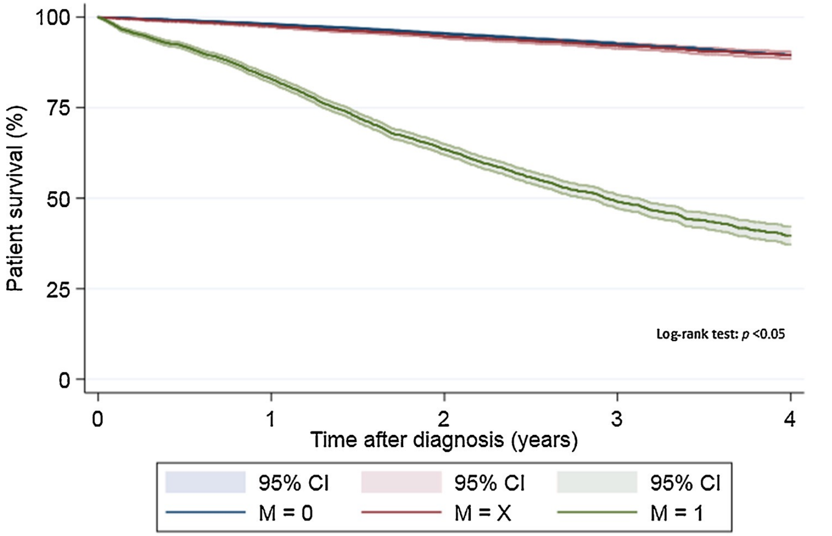
Contents lists available at ScienceDirect

Cancer Epidemiology

journal homepage: www.elsevier.com/locate/canep

Imputation of missing prostate cancer stage in English cancer registry data based on clinical assumptions

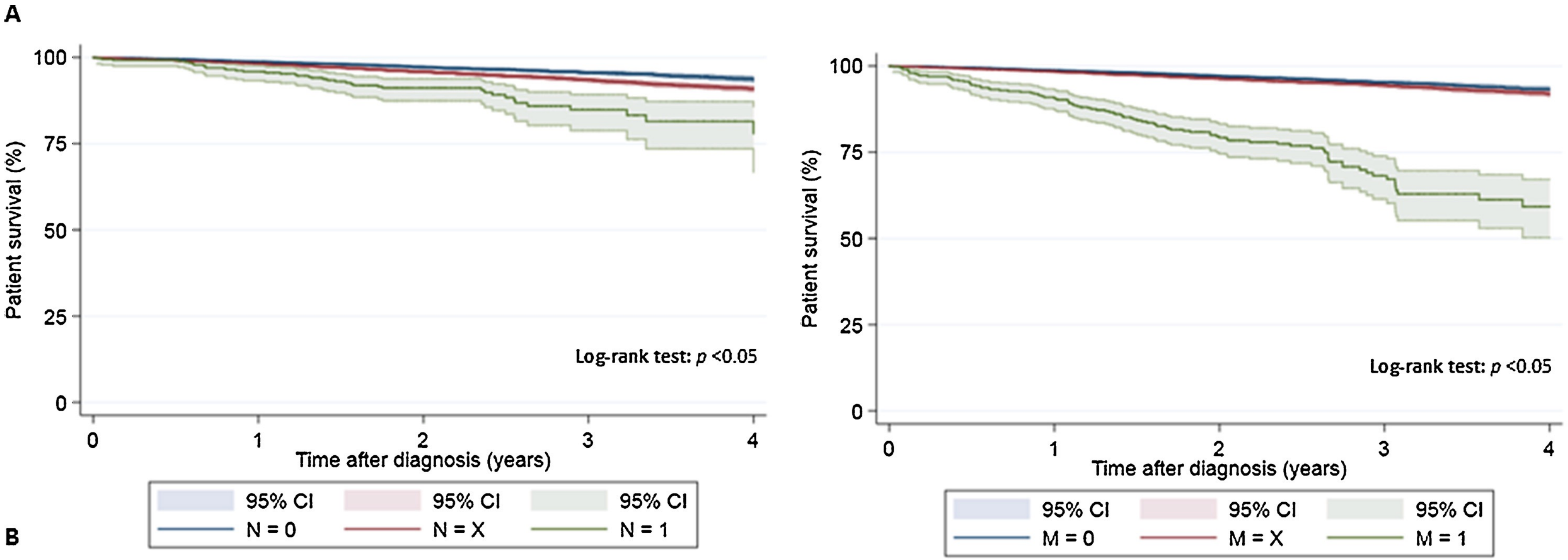
Matthew G. Parry[a,b,*], Arunan Sujenthiran[b], Thomas E. Cowling[a,b], Susan Charman[b], Julie Nossiter[a,b], Ajay Aggarwal[a,c,d], Noel W. Clarke[e,f], Heather Payne[g], Jan van der Meulen[a,b]



***Note:** Patient survival for men with M0 (blue line) or MX (red line) was very similar with narrow and overlapping confidence intervals (95%). Both lines therefore appear superimposed.

Overall survival for men with **complete N-stage (N1/N0)** showing the distribution of M-stage (M1/M0/missing M).

Cancer Epidemiol. 2019 Feb:58:44-51.

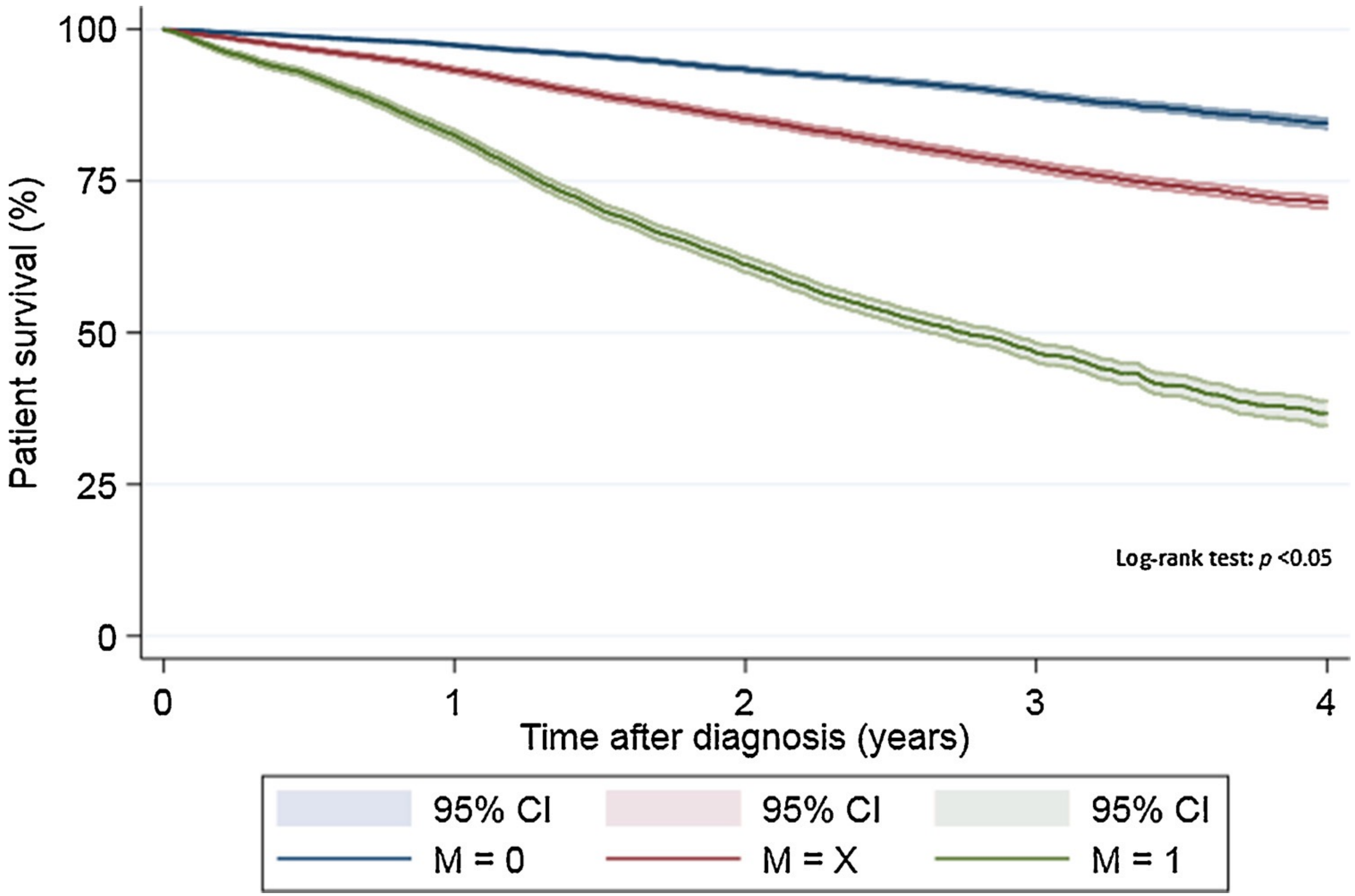


Overall survival for men with **low/intermediate-risk disease** (T1-2 and Gleason score ≤7) showing the distribution of: a. N-stage (N1/N0/missing N) b. M-stage (M1/M0/missing M)

3 clinical assumptions:

1. Recorded N-stage: missing M-stage → M0
2. Low/Intermediate-risk men: missing M-stage → M0
3. Low/Intermediate-risk men: missing N-stage → N0

Cancer Epidemiol. 2019 Feb:58:44-51.



Overall survival for men with **high-risk disease** (T3-4 or Gleason score ≥8) showing the distribution of M-stage (M1/M0/missing M).

Cancer Epidemiol. 2019 Feb:58:44-51.

2. Partial deterministic imputation + MI (PDI + MI)

- Increased the completeness of clinical staging
- Perform as well as multiple imputation
- More easily applicable for those without appropriate statistical software or expertise
- Less appropriate for use in cancer registries with less complete staging data

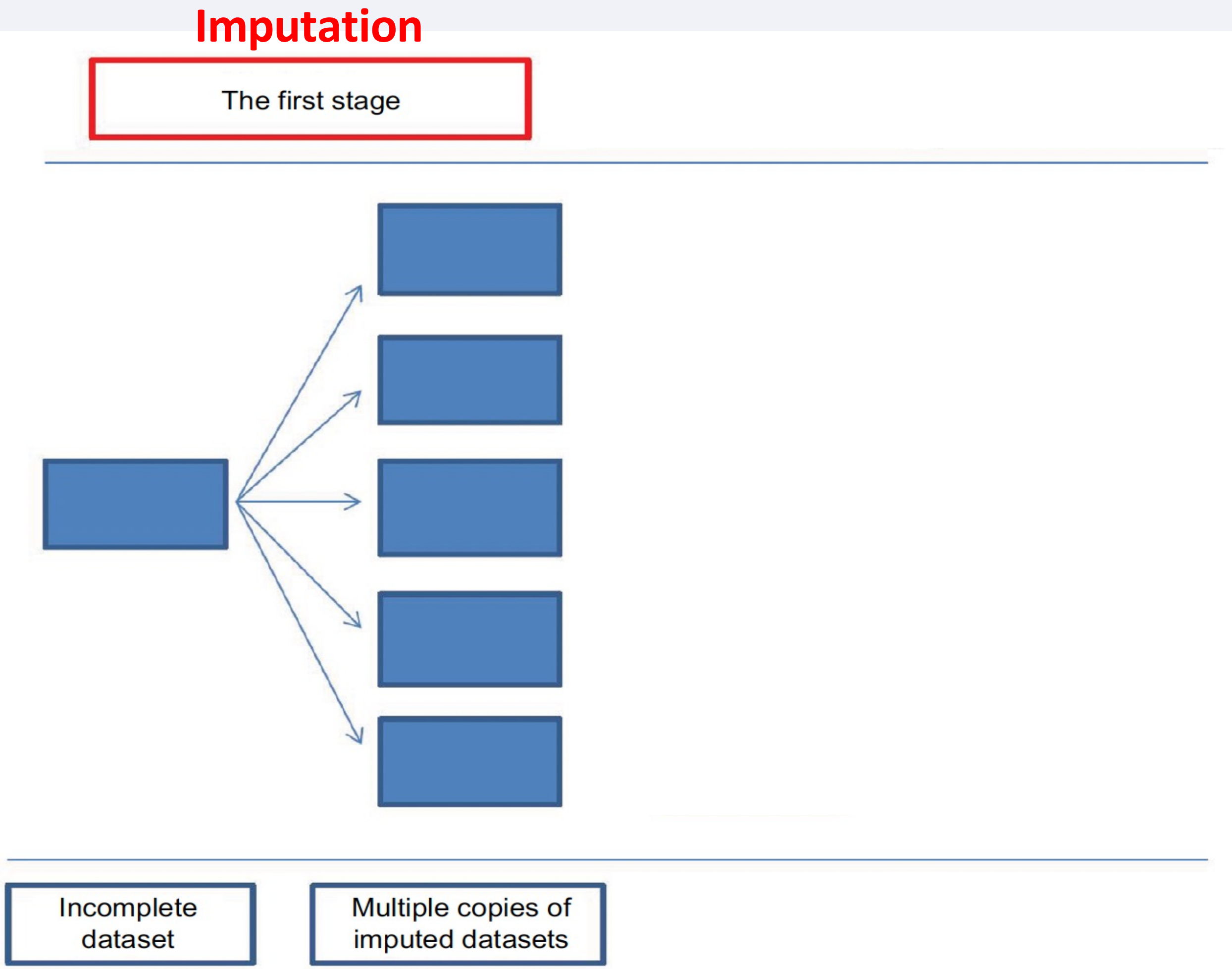
Cancer Epidemiol. 2019 Feb:58:44-51.

3. Standard MI

- All variables were included and missing data were imputed using MI.
- This model was identical to PDI + MI. The only difference was that M stage was not substituted to M0 prior to performing the multiple imputation procedure.

Journal of Clinical Epidemiology 155 (2023) 22-30

The 3 main stages of implementing MI



Imputation

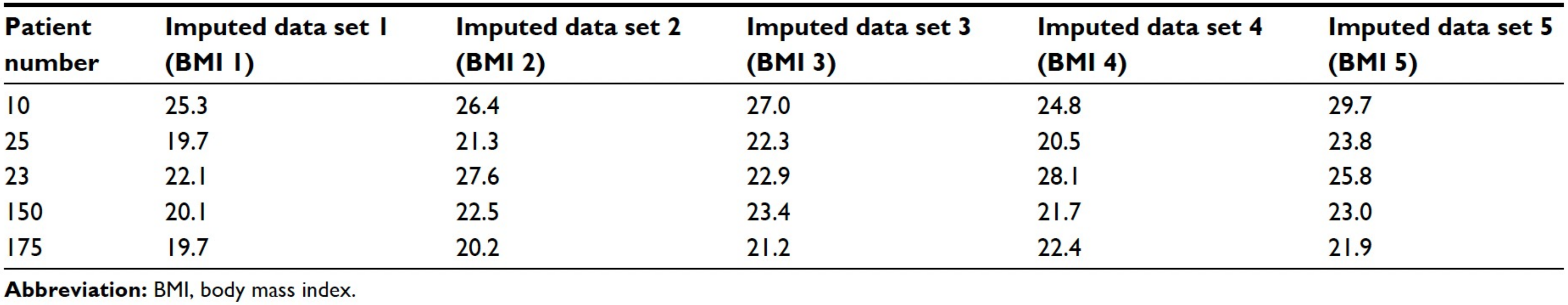
The first stage

Incomplete dataset

Multiple copies of imputed datasets

Clinical Epidemiology 2017:9 157–166

An example of the imputed missing BMI values generated with 5 imputed datasets

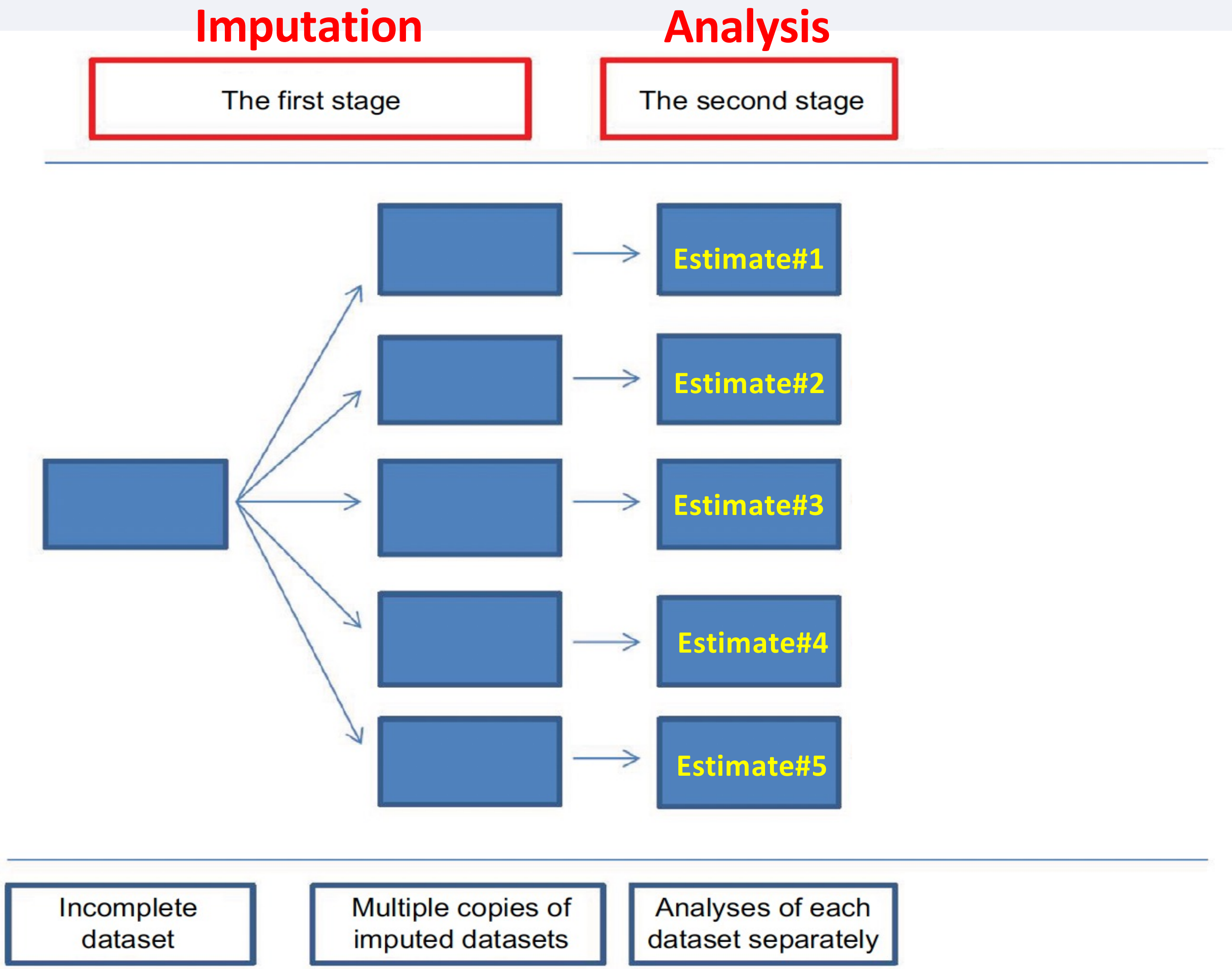


Patient number	Imputed data set 1 (BMI 1)	Imputed data set 2 (BMI 2)	Imputed data set 3 (BMI 3)	Imputed data set 4 (BMI 4)	Imputed data set 5 (BMI 5)
10	25.3	26.4	27.0	24.8	29.7
25	19.7	21.3	22.3	20.5	23.8
23	22.1	27.6	22.9	28.1	25.8
150	20.1	22.5	23.4	21.7	23.0
175	19.7	20.2	21.2	22.4	21.9

Abbreviation: BMI, body mass index.

Clinical Epidemiology 2017:9 157–166

The 3 main stages of implementing MI



Imputation

Analysis

The first stage

The second stage

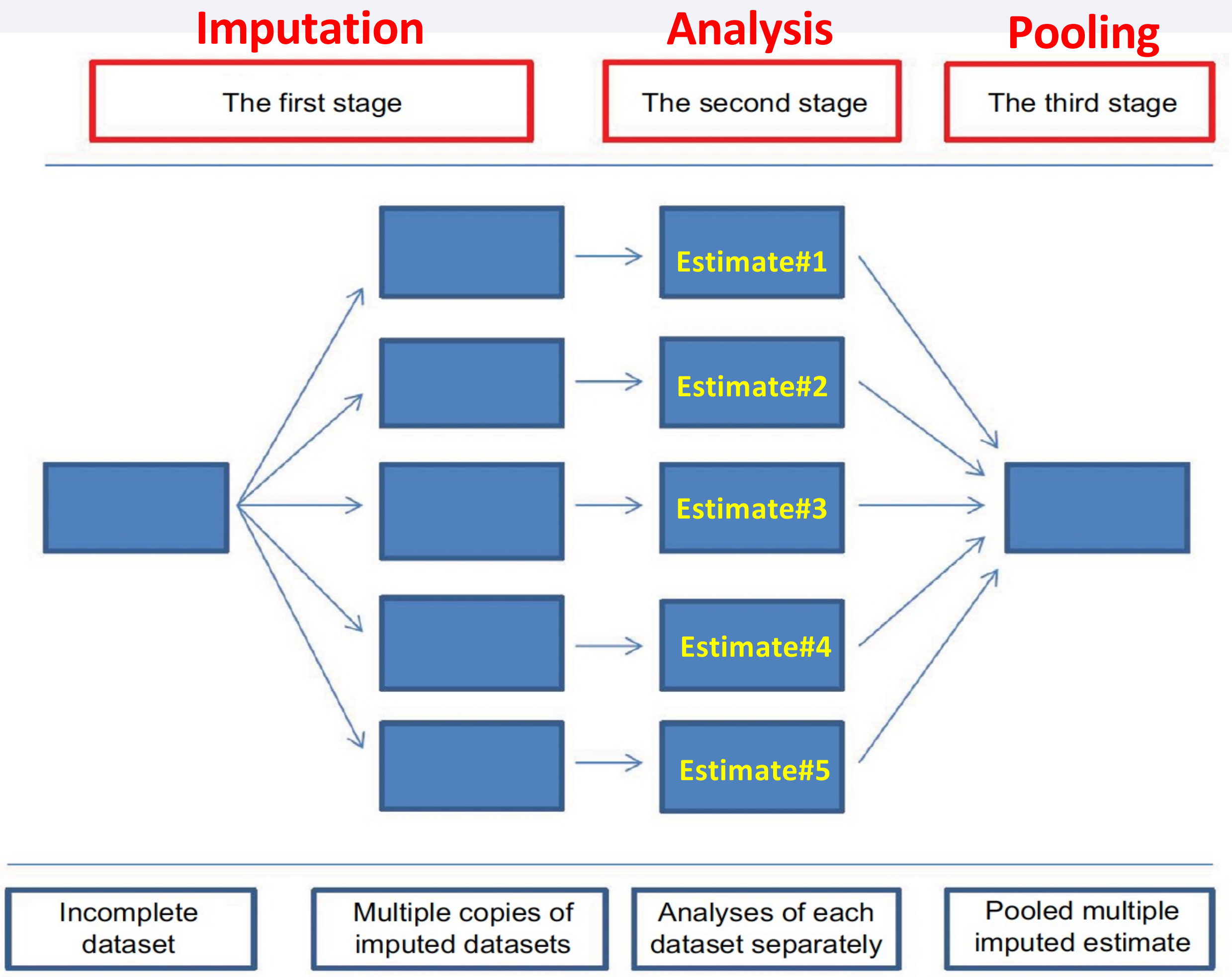
Incomplete dataset

Multiple copies of imputed datasets

Analyses of each dataset separately

Clinical Epidemiology 2017:9 157–166

The 3 main stages of implementing MI



Clinical Epidemiology 2017:9 157–166

Missing at random (MAR) assumption

- The validity of the MI methods relies on the plausibility of the MAR assumption
- T- tests and logistic regression analyses can be used to investigate if there is a relationship between variables with and without missing data
- MNAR can result in a large bias in estimates obtained after MI that operates under the MAR assumption

Journal of Clinical Epidemiology 155 (2023) 22e30

Oncology: Prostate/Testis/Penis/Urethra

Bias Due to Missing SEER Data in D'Amico Risk Stratification of Prostate Cancer

D'Amico staging requires all 3 variables

- Prostate specific antigen (PSA)
- T stage
- Gleason score

P: men with incident prostate cancer

E: patient age, race, Geographic region

O: *unclassified risk group* due to unknown variables

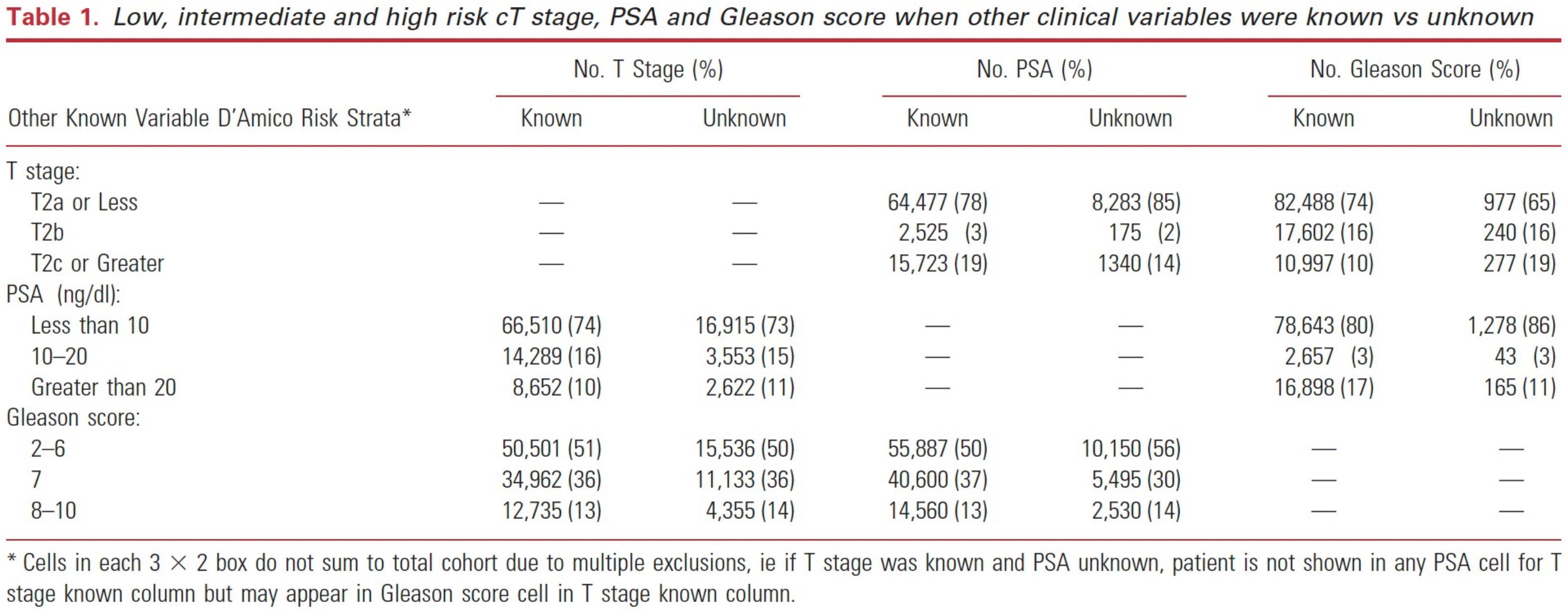
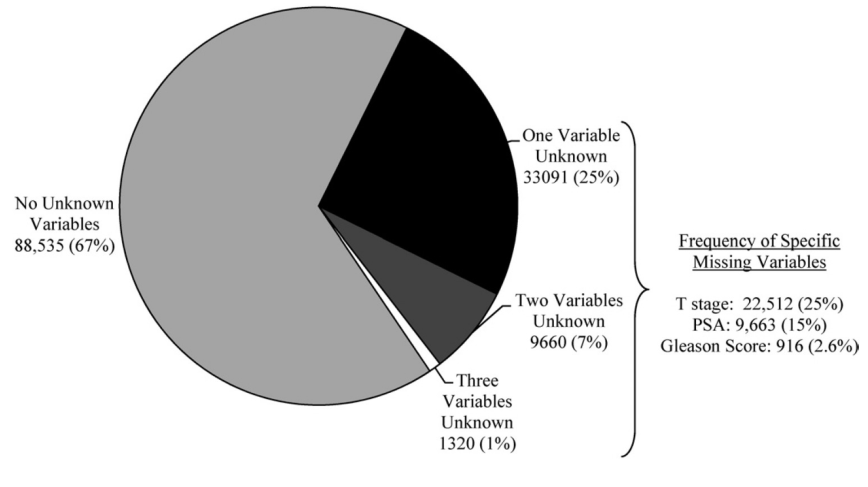


Table 1. *Low, intermediate and high risk cT stage, PSA and Gleason score when other clinical variables were known vs unknown*

	No. T Stage (%)		No. PSA (%)		No. Gleason Score (%)	
Other Known Variable D'Amico Risk Strata*	Known	Unknown	Known	Unknown	Known	Unknown
T stage:						
T2a or Less	—	—	64,477 (78)	8,283 (85)	82,488 (74)	977 (65)
T2b	—	—	2,525 (3)	175 (2)	17,602 (16)	240 (16)
T2c or Greater	—	—	15,723 (19)	1340 (14)	10,997 (10)	277 (19)
PSA (ng/dl):						
Less than 10	66,510 (74)	16,915 (73)	—	—	78,643 (80)	1,278 (86)
10–20	14,289 (16)	3,553 (15)	—	—	2,657 (3)	43 (3)
Greater than 20	8,652 (10)	2,622 (11)	—	—	16,898 (17)	165 (11)
Gleason score:						
2–6	50,501 (51)	15,536 (50)	55,887 (50)	10,150 (56)	—	—
7	34,962 (36)	11,133 (36)	40,600 (37)	5,495 (30)	—	—
8–10	12,735 (13)	4,355 (14)	14,560 (13)	2,530 (14)	—	—

* Cells in each 3 × 2 box do not sum to total cohort due to multiple exclusions, ie if T stage was known and PSA unknown, patient is not shown in any PSA cell for T stage known column but may appear in Gleason score cell in T stage known column.

J Urol 2012;187:2026e31.

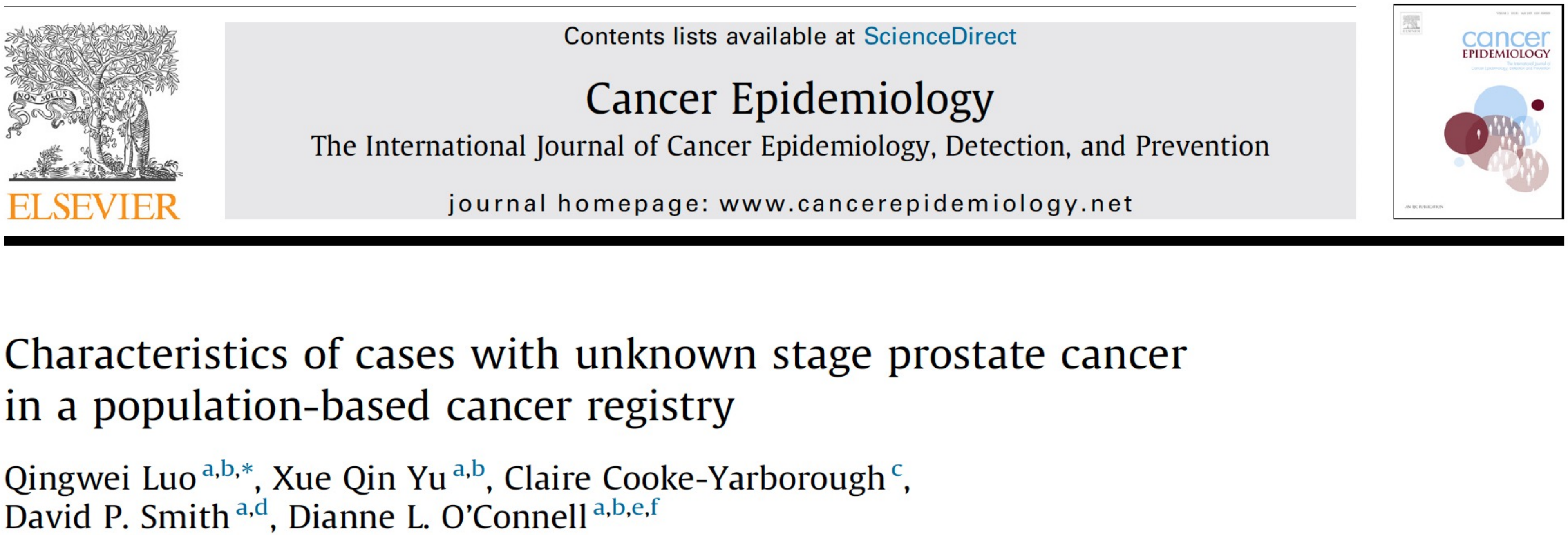


Frequency of unknown data on 132,606 patients

Table 2. *Unclassified DARG due to unknown variables by patient age and race*

	No. Pts (% unclassified DARG)	Unclassified DARG Adjusted OR (95% CI)	Probability Greater Than Chi-Square
Age:			
Less than 45	805 (32.2)	1.1 (1.0–1.3)	0.0801
45–54	11,757 (30.5)	1.1 (1.0–1.1)	0.0346
55–64	40,400 (29.3)	1.0 (referent)	—
65–74	48,066 (31.1)	1.1 (1.1–1.1)	<0.0001
75–84	27,135 (40.4)	1.5 (1.5–1.6)	<0.0001
85 or Greater	4,408 (55.8)	2.4 (2.3–2.6)	<0.0001
Unknown	35 (80)	6.6 (2.8–15.3)	<0.0001
Race:			
NonHispanic white	94,270 (33.2)	1.0 (referent)	—
NonHispanic black	15,093 (29.2)	0.8 (0.8–0.9)	<0.0001
Hispanic	11,722 (33.3)	1.1 (1.1–1.2)	<0.0001
Asian/Pacific Islanders	6,278 (27.1)	1.0 (0.9–1.0)	0.1188
AI/AN/other/unknown	5,243 (53.0)	2.4 (2.2–2.5)	<0.0001

J Urol 2012;187:2026e31.



ELSEVIER

Contents lists available at ScienceDirect

Cancer Epidemiology

The International Journal of Cancer Epidemiology, Detection, and Prevention

journal homepage: www.cancerepidemiology.net

Characteristics of cases with unknown stage prostate cancer in a population-based cancer registry

Qingwei Luo [a,b,*], Xue Qin Yu [a,b], Claire Cooke-Yarborough [c], David P. Smith [a,d], Dianne L. O'Connell [a,b,e,f]

P: Primary prostate cancer cases from New South Wales Central Cancer Registry (NSW CCR)

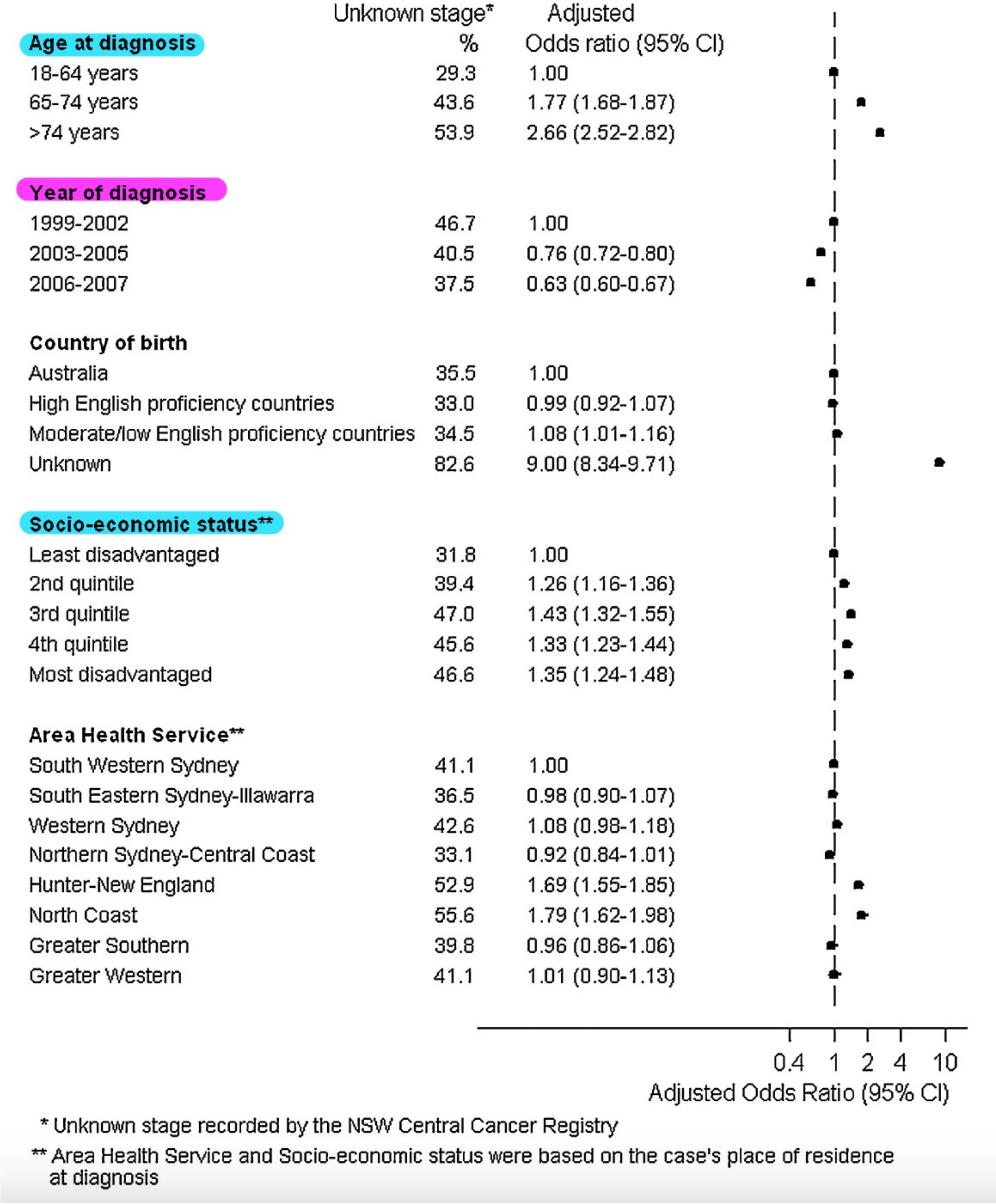
E: Patient characteristics

- age, place of residence at diagnosis, year of diagnosis and country of birth

O: Disease stage of prostate cancer

- localized, regional, distant or ‘‘unknown’’

Cancer Epidemiology 37 (2013) 813–819



	Unknown stage* %	Adjusted Odds ratio (95% CI)
Age at diagnosis		
18-64 years	29.3	1.00
65-74 years	43.6	1.77 (1.68-1.87)
>74 years	53.9	2.66 (2.52-2.82)
Year of diagnosis		
1999-2002	46.7	1.00
2003-2005	40.5	0.76 (0.72-0.80)
2006-2007	37.5	0.63 (0.60-0.67)
Country of birth		
Australia	35.5	1.00
High English proficiency countries	33.0	0.99 (0.92-1.07)
Moderate/low English proficiency countries	34.5	1.08 (1.01-1.16)
Unknown	82.6	9.00 (8.34-9.71)
Socio-economic status**		
Least disadvantaged	31.8	1.00
2nd quintile	39.4	1.26 (1.16-1.36)
3rd quintile	47.0	1.43 (1.32-1.55)
4th quintile	45.6	1.33 (1.23-1.44)
Most disadvantaged	46.6	1.35 (1.24-1.48)
Area Health Service**		
South Western Sydney	41.1	1.00
South Eastern Sydney-Illawarra	36.5	0.98 (0.90-1.07)
Western Sydney	42.6	1.08 (0.98-1.18)
Northern Sydney-Central Coast	33.1	0.92 (0.84-1.01)
Hunter-New England	52.9	1.69 (1.55-1.85)
North Coast	55.6	1.79 (1.62-1.98)
Greater Southern	39.8	0.96 (0.86-1.06)
Greater Western	41.1	1.01 (0.90-1.13)

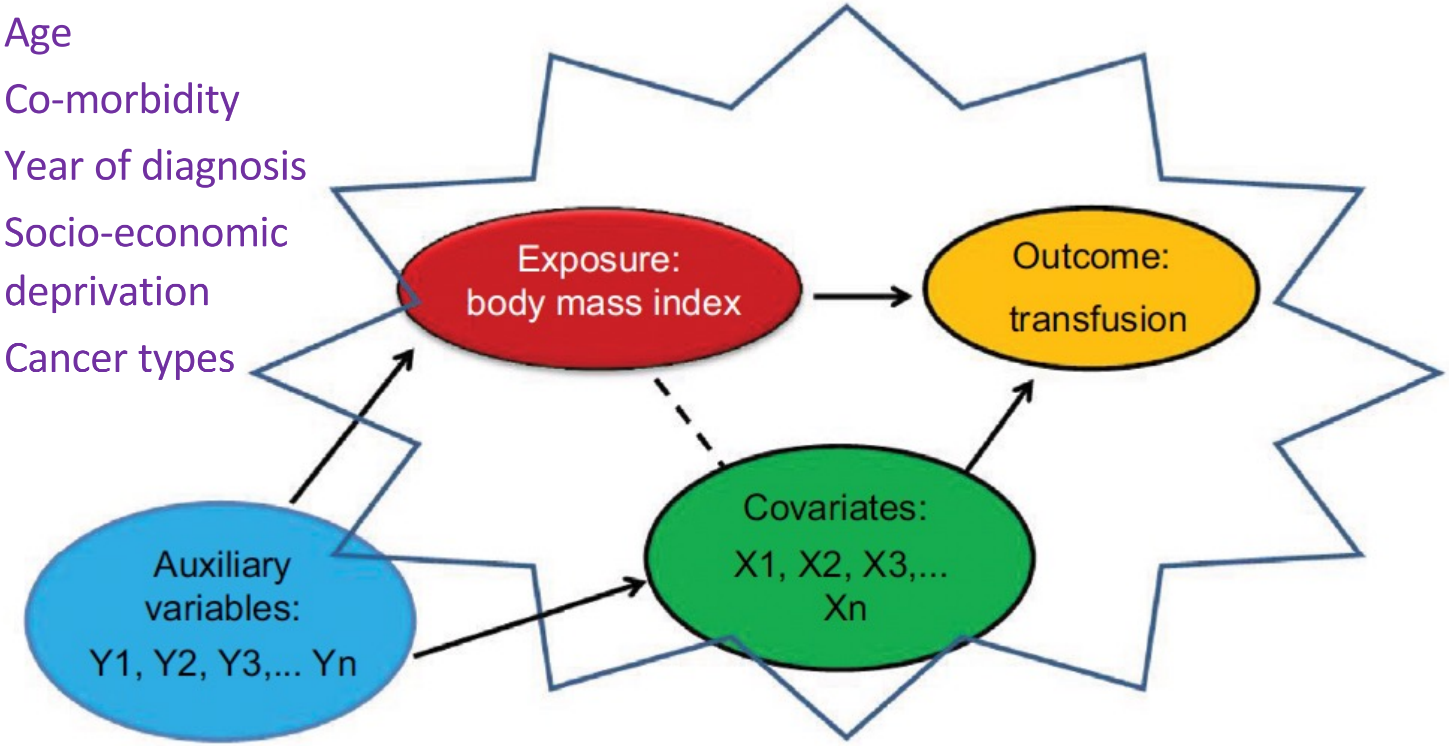
* Unknown stage recorded by the NSW Central Cancer Registry

** Area Health Service and Socio-economic status were based on the case's place of residence at diagnosis

Selection of variables

in order to create multiple imputed datasets when looking into the association BMI and transfusion risk.

- Age
- Co-morbidity
- Year of diagnosis
- Socio-economic deprivation
- Cancer types



Selection of variables to create multiple imputed datasets:

Auxiliary variables

The variables that are in the subsequent analysis model (exposure, covariates, and outcome)

variables within the original data that are not included in the analysis, but are correlated to the variables of interest or help to keep the missing process random

Clinical Epidemiology 2017:9 157–166

Which variables should be included in the multiple imputation model?

Auxiliary variables need to fulfill one of following criteria

1) The auxiliary variable should be associated with the values of the incomplete variables
2) The auxiliary variable should be associated with the value of the incomplete variables and the likelihood of the data being missing

Clinical Epidemiology 2017:9 157–166

Which variables should be included in the multiple imputation model?

If we are not sure, these relationships can be identified by setting up,

1) a logistic regression model with the missingness (as 0 or 1) being the outcome and auxiliary variables being the explanatory variables, or
2) a regression model with the incomplete variable as the outcome and auxiliary variables again as explanatory variables.

Clinical Epidemiology 2017:9 157–166

4. Restricted MI

- Only TNM stage, age, and year of diagnosis were included, and missing data were imputed using MI.
- In particular, PSA, Gleason score, survival time and indicator of cause of death (or censoring) were omitted from the imputation model.
- Survival data were included in a sensitivity analysis.

Journal of Clinical Epidemiology 155 (2023) 22-30

Multiple imputation example

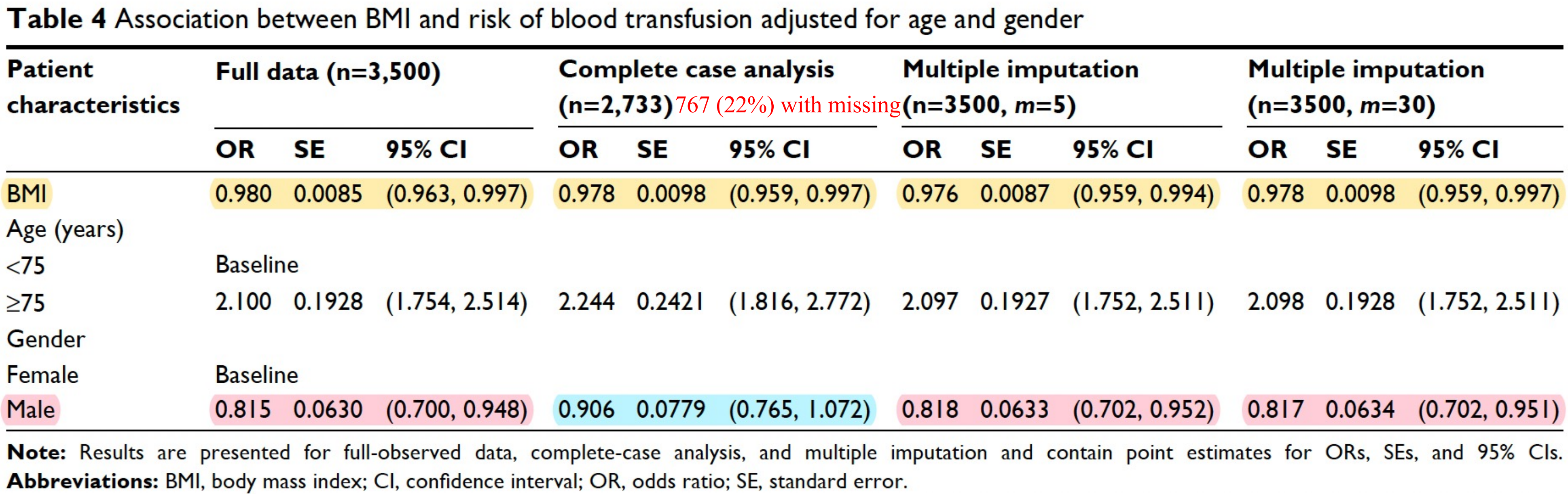
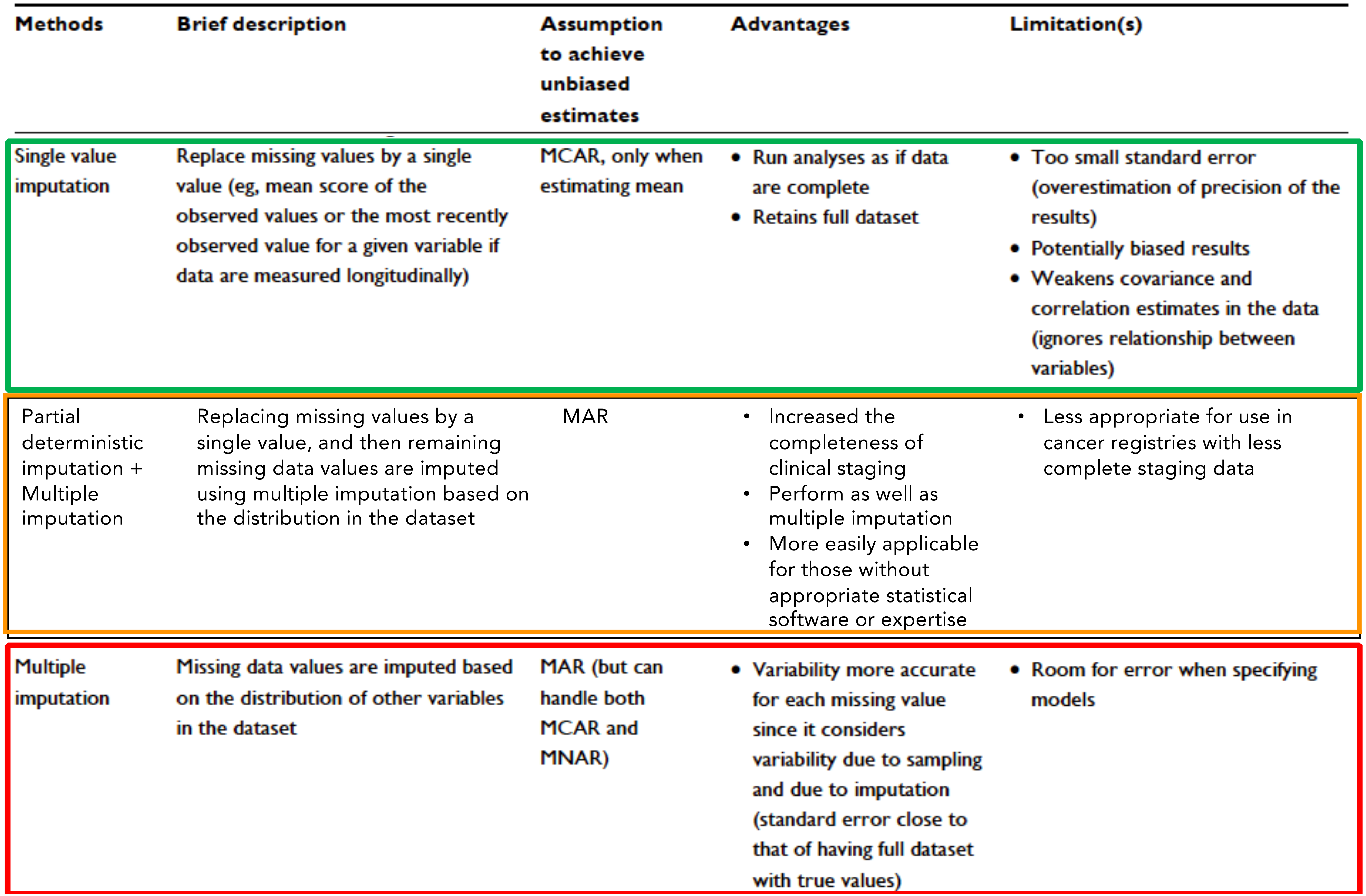


Table 4 Association between BMI and risk of blood transfusion adjusted for age and gender

Patient characteristics	**Full data (n=3,500)**			**Complete case analysis (n=2,733)** 767 (22%) with missing			**Multiple imputation (n=3500, *m*=5)**			**Multiple imputation (n=3500, *m*=30)**		
	OR	**SE**	**95% CI**	**OR**	**SE**	**95% CI**	**OR**	**SE**	**95% CI**	**OR**	**SE**	**95% CI**
BMI	0.980	0.0085	(0.963, 0.997)	0.978	0.0098	(0.959, 0.997)	0.976	0.0087	(0.959, 0.994)	0.978	0.0098	(0.959, 0.997)
Age (years)												
<75	Baseline											
≥75	2.100	0.1928	(1.754, 2.514)	2.244	0.2421	(1.816, 2.772)	2.097	0.1927	(1.752, 2.511)	2.098	0.1928	(1.752, 2.511)
Gender												
Female	Baseline											
Male	0.815	0.0630	(0.700, 0.948)	0.906	0.0779	(0.765, 1.072)	0.818	0.0633	(0.702, 0.952)	0.817	0.0634	(0.702, 0.951)

Note: Results are presented for full-observed data, complete-case analysis, and multiple imputation and contain point estimates for ORs, SEs, and 95% CIs.
Abbreviations: BMI, body mass index; CI, confidence interval; OR, odds ratio; SE, standard error.

Clinical Epidemiology 2017:9 157–166



Methods	Brief description	Assumption to achieve unbiased estimates	Advantages	Limitation(s)
Single value imputation	Replace missing values by a single value (eg, mean score of the observed values or the most recently observed value for a given variable if data are measured longitudinally)	MCAR, only when estimating mean	• Run analyses as if data are complete • Retains full dataset	• Too small standard error (overestimation of precision of the results) • Potentially biased results • Weakens covariance and correlation estimates in the data (ignores relationship between variables)
Partial deterministic imputation + Multiple imputation	Replacing missing values by a single value, and then remaining missing data values are imputed using multiple imputation based on the distribution in the dataset	MAR	• Increased the completeness of clinical staging • Perform as well as multiple imputation • More easily applicable for those without appropriate statistical software or expertise	• Less appropriate for use in cancer registries with less complete staging data
Multiple imputation	Missing data values are imputed based on the distribution of other variables in the dataset	MAR (but can handle both MCAR and MNAR)	• Variability more accurate for each missing value since it considers variability due to sampling and due to imputation (standard error close to that of having full dataset with true values)	• Room for error when specifying models

THANK
YOU
FOR
YOUR
ATTENTION