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

Abstract

Objectives: To study how handling missing data on M stage in a clinical cancer register affects estimates of incidence of metastatic prostate cancer.

Study Design and Setting: Estimates of age-standardized incidence of metastatic prostate cancer were obtained by the use of data in a population-based clinical cancer register in Sweden and using four methods for imputation of missing M stage. Adjusted survival was used to compare men with known and imputed M stage.

Results: The proportion of men with missing M stage was high (66%) and varied according to the risk group and over calendar time. The estimated incidence of metastatic disease varied depending on imputation method, with all methods indicating a decreasing incidence over time. A combination of deterministic imputation (DI) and multiple imputation (MI) produced adjusted survival curves for men with imputed M stage that best resembled the survival for men with known M stage.

Conclusions: Plausible estimates of incidence of metastatic prostate cancer in clinical cancer registers can be obtained by the use of a combination of DI of missing M stage and MI.

Keywords: Missing data; Prostate cancer; Incidence; Metastases; Imaging; TNM staging

1. Introduction

The incidence of de novo metastatic cancer (i.e., metastatic cancer at diagnosis) is an early proxy for cancer-specific mortality when evaluating interventions such as screening and easier and earlier access to health care. Incidence of de novo metastatic cancer is unaffected by new treatments, and it does not require a long observation time. Missing data on tumor node metastasis (TNM) variables is common [1,2] and temporal changes in use of imaging can influence the pattern of missingness in M stage. For example, efforts to discourage inappropriate use of bone imaging in men with low-risk prostate cancer in Sweden reduced the proportion of men with low-risk prostate cancer who underwent bone imaging from 45% in 1998 to 3% in 2009 [3]. Missing data may also vary over time

Conflicts of interest: The authors report no conflicts of interest.

Funding: This project was supported by the Swedish Cancer Society [19 00 30] and Region Uppsala and Uppsala University. Marcus Westerberg received financial support from the Center for Interdisciplinary Mathematics (CIM), Uppsala University. The sponsors had no involvement with the planning, execution, or completion of the study.

Disclaimer: Rolf Gedeborg is employed by the Medical Products Agency (MPA) in Sweden. The MPA is a Swedish Government Agency. The views expressed in this article may not represent the views of the MPA.

Author statement: M.W. contributed to conceptualization, methodology, software, formal analysis, investigation, project administration, visualization, validation, writing−original draft. K.B. contributed to conceptualization, writing−review and editing. R.G. contributed to conceptualization, methodology, writing−review and editing. S.I. contributed to conceptualization. writing−review and editing. L.H. contributed to conceptualization, writing−review and editing. H. G. contributed to conceptualization, methodology, project administration, data curation, writing−review and editing. Pär Stattin contributed to conceptualization, project administration, funding acquisition, supervision, resources, writing−review and editing.

* Corresponding author. Regional Cancer Center Midsweden, Uppsala University Hospital, 752 37, Uppsala, Sweden. Tel.: +46-18-617-71-00.
E-mail address: marcus.westerberg@uu.se (M. Westerberg).

https://doi.org/10.1016/j.jclinepi.2022.12.008

What is new?

Key findings

- The estimated incidence of metastatic prostate cancer (M1) varied depending on how missing data on M stage was handled. Simply substituting all missing M stage with M0 underestimated the incidence of M1. Deterministic imputation of missing M stage to M0 among men with low risk of metastases in combination with multiple imputation yielded similar survival comparing men with known and missing M-stage.

What this adds to what is known?

- Information on serum of prostate-specific antigen levels and Gleason score, in addition to tumor node metastasis stage is necessary to yield plausible estimates when imputing missing M stage.

What is the implication and what should change now?

- Missing M stage cannot by default be deterministically imputed to M0. Analyses of incidence trends of metastatic prostate cancer should be complemented with sensitivity analyses and information on how missing M stage was handled.

due to revised coding principles in cancer staging systems. An example is the removal of the category "Mx" for unknown metastatic status in the seventh edition of the TNM classification, with the result that men who have not undergone bone imaging are now classified as M0 [4,5]. Trends in the incidence of de novo metastatic cancer may be biased unless missing M stage is handled appropriately because the reasons for missing M stage vary over calendar time and across risk categories [6−8].

1.1. Aim of the study

The aim of the study was to assess statistical methods for estimating the age-standardized incidence of de novo metastatic prostate cancer when M stage is missing for a large proportion of men. The methods used should account for missing data that vary over calendar time and are related to other measured and unmeasured clinical variables.

2. Materials

All men diagnosed with prostate cancer from 2000 to 2019 registered in the National Prostate Cancer Register (NPCR) of Sweden were included [9]. The NPCR includes data on diagnostic work-up, tumor characteristics, and primary treatment. Data linkages in the Prostate Cancer data Base Sweden (PCBaSe) were performed between the NPCR, the National Patient Register, the National Cancer Register, and the National Cause of Death Register by use of the unique Swedish personal identity number [10]. The following variables were extracted from PCBaSe: age at diagnosis, year of diagnosis, serum level of prostate-specific antigen (PSA), clinical TNM stage, Gleason score (GS) of the diagnostic biopsy cores or World Health Organization (WHO) grade in fine needle biopsies, mode of detection (lower urinary tract symptoms, other symptoms, and asymptomatic), primary treatment, Charlson Comorbidity Index (CCI), survival time, and status (cause of death [prostate cancer or other causes] or censoring). Follow-up ended at the time of death or at the end of follow-up (December 31, 2019). Primary treatment was categorized into radical treatment (radical prostatectomy or radiotherapy), androgen deprivation therapy (ADT) (gonadotropin-releasing hormone, antiandrogens [bicalutamide] or orchidectomy), deferred treatment (active surveillance or watchful waiting) and other or unknown treatment (other). The CCI was based on discharge diagnoses, excluding prostate cancer and metastases, from the National Patient Register up to 10 years prior to prostate cancer diagnosis. Data on all men alive each year between the ages of 40 and 100 years were obtained from Statistics Sweden (SCB) [11].

Men with prostate cancer were categorized according to the risk of metastatic disease at diagnosis:

Low metastatic risk: PSA <20 ng/mL, T1-2, and GS ≤ 7 or WHO grade 1−2 if GS is missing,

High metastatic risk: PSA $\geq$ 20 ng/mL, T3-4, GS >7, or WHO grade 3 if GS is missing,

Unknown metastatic risk: if missing any of PSA, T stage, and simultaneously both of GS and WHO grade.

The categorization was designed to closely match the current Swedish clinical guidelines for use of imaging in the diagnostic workup of men with prostate cancer [12].

3. Methods

We estimated the age-standardized incidence of de novo metastatic prostate cancer according to the age distribution in Sweden 2000 by using direct standardization [13]. To obtain an annual estimate of the proportion of M1 among all men alive in each age strata in the presence of missing data on M stage we used four different methods based on deterministic imputation (DI) and multiple imputation (MI) using the R package mice [14,15] as described below. The number of MIs was set to 128 [16]. The definition of M stage used prior to 2011 was recreated for the whole cohort; i.e., M stage was considered missing if the man had not undergone imaging to assess metastatic status. Adjusted survival curves stratified by M stage were used to compare known and imputed M stage among men with M0 and M1, respectively, and these were obtained by the method

of weighting to account for potential differences in baseline characteristics [17]. See Supplementary Materials for further details on the methods, specification of the imputation models, weight diagnostics, and sensitivity analyses.

3.1. Deterministic imputation

M stage was substituted to M0 for all men with missing M stage. This corresponds to a situation where only positive imaging results are registered and imaged men with M0 cannot be differentiated from nonimaged men, as in the current Union for International Cancer Control classification [5].

3.2. Partial deterministic imputation + multiple imputation

For men with low-risk prostate cancer [18] the National Swedish guidelines for prostate cancer recommend against imaging as the prevalence of M1 among these men is very low [3]. M stage was therefore first substituted to M0 for all men categorized as low metastatic risk with missing M stage, and then remaining missing data in M stage and all other variables (e.g., PSA and N stage) was imputed using MI including all variables listed in the Materials section.

3.3. Standard MI

All variables listed in the Material section were included and missing data were imputed using MI. This method corresponds to a standard implementation of MI without any prior deterministic imputation.

3.4. Restricted MI

Many registers contain a limited number of variables used in clinical practice, such as the National Cancer Registry in Sweden that only registers TNM and no other clinical variables or survival data. To simulate this scenario only TNM stage, age, and year of diagnosis were included, and missing data were imputed using MI. Survival data were included in a sensitivity analysis, see Supplementary Materials.

4. Results

4.1. Baseline characteristics

There were 190,420 men diagnosed with prostate cancer between 2000 and 2019 in NPCR. Baseline characteristics by M stage are summarized in Table 1. Of which 126,102 men (66%) had missing M stage; 15,526 men (8%) were M1, constituting 24% of all imaged men. Men with missing M stage had similar characteristics as men with M0 with respect to age at diagnosis, CCI, and mode of detection. The PSA, T stage and GS, however, indicated more favorable disease characteristics in men with missing M stage. Thirty six percent of men with M0 and 3% of men with M1 were categorized as low metastatic risk. The corresponding proportion for men with missing M stage was 70%, and these were substituted to M0 by the DI method and the partial deterministic imputation + MI (PDI + MI) method prior to MI.

The annual number of men diagnosed with prostate cancer increased during the study period, while the annual number of men categorized as high metastatic risk was stable in all age groups. Simultaneously, the proportion of imaged men (i.e., known M stage) decreased from 48% in 2000 to 23% in 2008. This was followed by an increase to 37% in 2019 which was most pronounced among men aged 70 years or above categorized as high metastatic risk (Supplementary Figure 1).

4.2. Baseline characteristics after imputation

The proportions of men with imputed M1 among men with missing M stage were 7%, 10%, and 16% when applying PDI + MI, standard MI (SMI), and restricted MI (RMI), respectively. Among men with imputed M1, the proportion categorized as low metastatic risk varied substantially (1−40%) depending on the imputation method used (Supplementary Table 1), compared with 4% among men with known M1 (Table 1). When using PDI + MI, men with imputed M1 were older, had higher CCI, fewer were detected through a health checkup, and most men were assigned to primary treatment by ADT compared to other methods for imputation. The tumor characteristics among men with imputed M0 were similar across methods and tended toward more favorable disease characteristics compared to men with known M0 (Supplementary Table 1).

4.3. Incidence of metastatic prostate cancer

The estimated age-standardized incidence of de novo metastatic prostate cancer varied markedly between the four applied methods (Figure 1). The estimated incidences were 43, 70, 74, and 91 per 100,000 men in 2000 for each method DI, PDI + MI, SMI, and RMI, respectively. Both the estimated incidences, as well as the difference in estimated incidences between methods, decreased with time and were 32, 40, 50, and 57 per 100,000 men for DI, PDI + MI, SMI, and RMI, respectively in 2019. However, the estimated incidence curve was u-shaped for DI with a minimum of 26 per 100,000 men in 2012.

The estimated annual incidence of men with de novo metastatic prostate cancer categorized as low metastatic risk varied between methods (Supplementary Figure 2); SMI initially yielded a decrease followed by an increase over time, from 5 in 2000 to 11 per 100,000 men in 2019, and RMI yielded an increase over time, from 12 to 17 per 100,000 men, whereas DI and PDI + MI were stable around 2 per 100,000 men. The estimated annual incidence

Table 1. Baseline characteristics of men diagnosed with prostate cancer in PCBaSe between 2000 and 2019 by M stage determined by imaging. Men that did not undergo bone imaging have missing M stage. Column-wise percentages are indicated with () and row-wise percentages are indicated with []

	All		M Stage M0		M Stage M1		Missing M stage	
■	*n*	(%)	*n*	(%) [%]	*n*	(%) [%]	*n*	(%) [%]
N	190,420	(100)	48,792	(100) [26]	15,526	(100) [8]	126,102	(100) [66]
Age at diagnosis, yr								
<60	23,851	(13)	5,329	(11) [22]	1,027	(7) [4]	17,495	(14) [73]
60–69	70,888	(37)	18,480	(38) [26]	3,900	(25) [6]	48,508	(38) [68]
70–74	36,729	(19)	11,200	(23) [30]	3,050	(20) [8]	22,479	(18) [61]
75–80	28,945	(15)	7,973	(16) [28]	3,140	(20) [11]	17,832	(14) [62]
80+	30,007	(16)	5,810	(12) [19]	4,409	(28) [15]	19,788	(16) [66]
Year of diagnosis								
2,000–2,005	50,744	(27)	14,894	(31) [29]	4,955	(32) [10]	30,895	(25) [61]
2,006–2,011	56,868	(30)	10,005	(21) [18]	3,527	(23) [6]	43,336	(34) [76]
2,012–2,019	82,808	(43)	23,893	(49) [29]	7,044	(45) [9]	51,871	(41) [63]
Charlson Comorbidity Index								
0	137,465	(72)	35,967	(74) [26]	9,705	(63) [7]	91,793	(73) [67]
1	25,091	(13)	6,498	(13) [26]	2,640	(17) [11]	15,953	(13) [64]
2	16,853	(9)	3,933	(8) [23]	1,769	(11) [10]	11,151	(9) [66]
3+	11,011	(6)	2,394	(5) [22]	1,412	(9) [13]	7,205	(6) [65]
PSA (ng/mL)								
Median (Q1, Q3)	10 (6-24)		15 (8-30)		138 (39-503)		8 (5-14)	
0–9	94,545	(50)	16,362	(34) [17]	1,038	(7) [1]	77,145	(61) [82]
10–19	37,144	(20)	12,975	(27) [35]	1,140	(7) [3]	23,029	(18) [62]
20–49	25,953	(14)	12,019	(25) [46]	2,315	(15) [9]	11,619	(9) [45]
50–99	10,975	(6)	4,140	(8) [38]	2,128	(14) [19]	4,707	(4) [43]
100–499	11,288	(6)	2,554	(5) [23]	4,705	(30) [42]	4,029	(3) [36]
500+	5,974	(3)	314	(1) [5]	4,028	(26) [67]	1,632	(1) [27]
Missing	4,541	(2)	428	(1) [9]	172	(1) [4]	3,941	(3) [87]
T stage								
1	89,350	(47)	16,343	(33) [18]	1,261	(8) [1]	71,746	(57) [80]
2	57,496	(30)	19,043	(39) [33]	3,290	(21) [6]	35,163	(28) [61]
3	32,854	(17)	11,725	(24) [36]	7,497	(48) [23]	13,632	(11) [41]
4	5,986	(3)	879	(2) [15]	2,859	(18) [48]	2,248	(2) [38]
Missing	4,734	(2)	802	(2) [17]	619	(4) [13]	3,313	(3) [70]
N stage								
0	39,849	(21)	21,544	(44) [54]	1,867	(12) [5]	16,438	(13) [41]
1	6,522	(3)	2,986	(6) [46]	2,496	(16) [38]	1,040	(1) [16]
Missing	144,049	(76)	24,262	(50) [17]	11,163	(72) [8]	108,624	(86) [75]
Gleason sum or WHO grade								
GS 6/WHO grade 1	76,341	(40)	10,397	(21) [14]	780	(5) [1]	65,164	(52) [85]
GS 7/WHO grade 2	69,224	(36)	21,700	(44) [31]	3,930	(25) [6]	43,594	(35) [63]
GS 8-10/WHO grade 3	40,496	(21)	16,216	(33) [40]	9,670	(62) [24]	14,610	(12) [36]
Missing[a]	4,359	(2)	479	(1) [11]	1,146	(7) [26]	2,734	(2) [63]
Metastatic risk								
Low metastatic risk	105,952	(56)	17,387	(36) [16]	536	(3) [1]	88,029	(70) [83]
High metastatic risk	73,378	(39)	29,878	(61) [41]	13,455	(87) [18]	30,045	(24) [41]
Unknown metastatic risk	11,090	(6)	1,527	(3) [14]	1,535	(10) [14]	8,028	(6) [72]
Mode of detection								
Health check-up	76,891	(40)	20,186	(41) [26]	2,381	(15) [3]	54,324	(43) [71]
Lower urinary tract symptoms	56,613	(30)	14,286	(29) [25]	4,404	(28) [8]	37,923	(30) [67]

(*Continued*)

Table 1. Continued

	All		M Stage M0		M Stage M1		Missing M stage	
■	*n*	(%)	*n*	(%) [%]	*n*	(%) [%]	*n*	(%) [%]
Other symptoms	50,268	(26)	12,754	(26) [25]	8,346	(54) [17]	29,168	(23) [58]
Missing	6,648	(3)	1,566	(3) [24]	395	(3) [6]	4,687	(4) [71]
Primary treatment								
Radical treatment[b]	74,752	(39)	28,828	(59) [39]	364	(2) [0]	45,560	(36) [61]
Androgen deprivation therapy	52,378	(28)	13,076	(27) [25]	14,392	(93) [27]	24,910	(20) [48]
Deferred treatment[c]	54,809	(29)	5,426	(11) [10]	261	(2) [0]	49,122	(39) [90]
Other	8,481	(4)	1,462	(3) [17]	509	(3) [6]	6,510	(5) [77]
Follow-up and status								
Median follow-up (Q1, Q3)	6 (3-10)		6 (3-10)		2 (1-4)		6 (3-10)	
Censored	119,272	(63)	32,068	(66) [27]	3,807	(25) [3]	83,397	(66) [70]
Death by prostate cancer	29,358	(15)	6,708	(14) [23]	9,101	(59) [31]	13,549	(11) [46]
Death by other causes	41,790	(22)	10,016	(21) [24]	2,618	(17) [6]	29,156	(23) [70]

PCBaSe, prostate cancer data base sweden; PSA, prostate-specific antigen; WHO, world health organization; GS, Gleason score.
[a] If GS is missing then WHO grade is reported if known.
[b] Radical treatment includes radical prostatectomy and radical radiotherapy.
[c] Deferred treatment includes active surveillance and watchful waiting.

of men with de novo metastatic prostate cancer categorized as high metastatic risk was similar for all methods except DI (Supplementary Figure 3).

4.4. Survival

The adjusted 5-year overall survival curves for men with known M0 or M1, and for men with missing M stage imputed as M0 or M1 are shown in Figure 2. When applying the methods PDI + MI and SMI, the survival curves for men with imputed M stage closely matched those for men with known M stage when considering all men and men categorized as high metastatic risk. Among men categorized as low metastatic risk, the number of imputed M1 according to PDI + MI were few ($n = 98$), making any comparison of survival uncertain. The adjusted survival curves for men with known and imputed M1 categorized as low metastatic risk separated immediately when applying the SMI method, and the RMI method yielded survival curves that did not match particularly well in any of the strata. The results were similar for prostate cancer specific survival (Supplementary Figure 4) and in unadjusted analyses (Supplementary Figures 5 and 6).

5. Discussion

5.1. Summary of findings

The estimated age-standardized incidence of de novo metastatic prostate cancer differed markedly between the

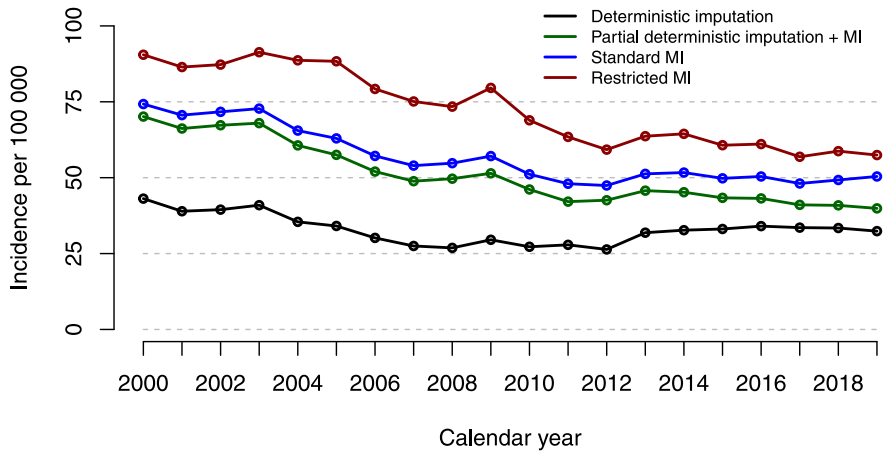

Fig. 1. Standardized incidence of metastatic prostate cancer (with respect to the age distribution in 2000) by statistical method. M stage was imputed for men with missing M stage, i.e., men that did not undergo bone imaging. Data on use of bone imaging were not registered in NPCR for men diagnosed in 2011. Men registered as M1 in 2011 with no information on imaging were considered as M1 when using DI. MI, multiple imputation; DI, deterministic imputation; PDI + MI, partially deterministic imputation + MI; SMI, standard MI; RMI, restricted MI; NPCR, National prostate cancer register of Sweden.

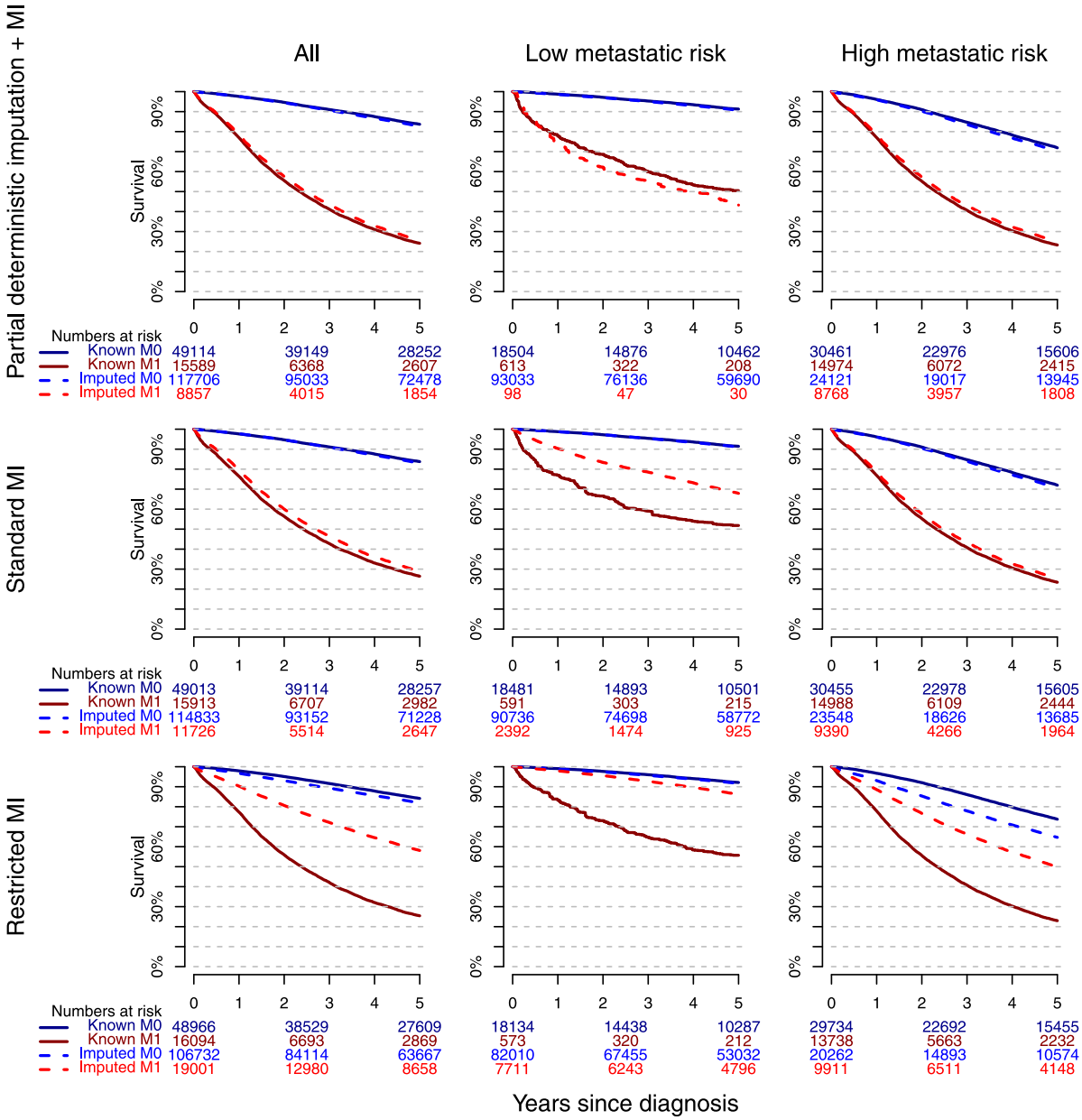

Fig. 2. Adjusted overall survival averaged over the multiple imputations, for all men and stratified by those with low metastatic risk and high metastatic risk. M stage was imputed for men with missing M stage, i.e., men that did not undergo bone imaging. The strata defined by M stage therefore vary across imputations and imputation methods. The survival estimates for known M stage were not based on a complete-case analysis because the weights were computed based on both observed and imputed data. Numbers at risk, reported as averages over the multiple imputations, were computed in the weighted population, and may therefore be different between the imputation methods for men with known M stage. Some men had missing data in PSA or GS and could not be categorized into low or high metastatic risk even after imputation using the restricted MI (that omitted PSA and GS). PSA, prostate-specific antigen; GS, Gleason score.

methods used to handle missing data in metastatic status. Partial deterministic imputation + multiple imputation simultaneously yielded a small number of men with imputed M1 among men with low metastatic risk and a survival of imputed M stage that best resembled that of observed M stage.

5.2. Validity of different methods for imputation of M stage

Deterministic imputation likely underestimates the incidence of M1, which mostly depends on the changing use of imaging over calendar time among men older than 70 years with high metastatic risk. Randomized clinical trials have

shown that radical radiotherapy with neoadjuvant and adjuvant ADT increase survival in men with locally advanced prostate cancer [19,20], which likely has led to a more comprehensive workup of men with high metastatic risk in more recent years. This likely explains both the increase of imaging in these men after 2008 and the U-shape of the incidence curve.

The validity of the MI methods relies on the plausibility of the missing at random (MAR) assumption [14]. It is recommended to include as many auxiliary variables as possible in the analysis to increase the plausibility of MAR [21,22], since such variables may explain systematic differences between those with observed and missing data. When such variables are not available or omitted, data can no longer be considered MAR and is instead missing not at random (MNAR) [14]. In this study, missing information on variables that predict the risk of metastases and the probability of undergoing imaging was considered the primary reason why data could be MNAR. MNAR can result in a large bias in estimates obtained after MI that operates under the MAR assumption.

Using subject matter knowledge is crucial when data are missing frequently and missingness may be MNAR. Based on recommendations in guidelines on the use of imaging, we hypothesized that men with baseline cancer characteristics indicating a low risk for metastatic disease and who did not undergo imaging were unlikely to have metastases. Substituting missing M stage with M0 for these men likely results in a negligible underestimation of M1 disease. We did not expect systematic differences between imaged and nonimaged men with high metastatic risk on the risk of metastases. This motivated the use of the PDI + MI method. The PDI + MI produced the most convincing imputations among the considered methods based on the low number of men with imputed M1 and low metastatic risk and on the similarity of the survival curves. However, the validity of estimated incidence based on this method depends on how well it approximates the truth, which is unknown, and we were unable to test the above assumptions. Therefore, the findings do not prove that the method is valid. Ideally, a validation study should be performed where a random selection of cases with missing M stage was subjected to a patient record review to try to determine M stage and/or the reason for missingness.

Restricted MI did not include survival time or cause of death in the imputation model and did not produce similar adjusted survival curves when comparing men with known and imputed M stage and was thus unable to adequately impute M stage, particularly among men with low metastatic risk. Consequently the annual incidence of metastatic prostate cancer was likely overestimated with this method.

5.3. Other studies

Other studies have reported an association between missing data in stage and comorbidity [23] and age [24−26]. Missing data in prostate cancer stage in the English Cancer Registry were imputed using a combination of substitution (deterministic imputation) followed by MI [2]. The authors observed an increase from 6% to 8% in the proportion of known metastatic prostate cancer between 2010 and 2013, which could potentially be due to changes in use of imaging as suggested by the large decrease over time in the proportion of men missing cancer stage, from 83.1% to 32.5%.

In an Australian cohort the validity of MI for missing cancer stage at diagnosis was assessed by cross-linkage with data from health care records [27]. The authors concluded that MI may be an appropriate method to handle missing data on cancer stage in a cancer registry particularly when more clinical variables were available. However any differences in clinical practice (e.g., diagnostic routines and use of imaging, which was not reported) and data registration (e.g., only summary stage was available and not separate TNM stage) makes it difficult to assess whether their findings are applicable in our study.

5.4. Implications for data registration and coding

Our results indicate that it is instrumental to have access to data on use of imaging to determine which men had known M stage, or else one cannot assess the potential magnitude of the underestimation of incidence. Such data may not be available for example if the M classification applied does not include a category ‘‘Mx’’ that indicates that imaging was not performed and unknown M stage is coded as M0, and if there is no other variable indicating whether imaging was performed. It is also important to be able to distinguish between whether M stage was determined by imaging or if men were coded M0 if there is no obvious signs of metastasis [4] and M1 if PSA > 100 ng/mL when imaging results were not reported [28,29].

5.5. Strengths and limitations

Data quality in NPCR has been shown to be high [30]. An important strength was the availability of several auxiliary variables, most with negligible amount of missing data, which predict M stage and missingness in M stage. This increased the plausibility of the MAR assumption. By comparing results of the imputation methods for missing M stage we gained insights into how data availability and handling of missingness in M stage may affect incidence estimates.

Limitations of our study include the large proportion of missing data in M stage (66%) and missing data are predictors for imputing M stage (e.g., 75.6% for N stage) that may affect the performance of MI [31]. Methods such as single-photon emission computerized tomography and positron emission tomography have higher sensitivity and specificity [32,33] than bone scintigraphy and changes in use of imaging modalities over time can cause bias. We were unable to

assess this potential bias due to lack of such data. Moreover any temporal changes in assessment and definition of the auxiliary variables may also be a source of bias. For example, the Gleason classification has been modified during the study period [34−37].

Our study focused on analyses of register data for epidemiological, population-level studies, and the concepts and implications of this article apply to statistical aspects of missing data and are not intended to be adopted for clinical practice. However the results can help guide instruction for coding in cancer registries and clinical databases.

6. Conclusions

The amount of missing data in metastatic status is often high even in clinical cancer registers with otherwise comprehensive data and the estimated age-standardized incidence of de novo metastatic prostate cancer is sensitive to how missing data in metastatic status is handled. Substituting missing M stage with M0 underestimates the incidence. The most convincing results were obtained from imputations of missing M stage using DI of missing M stage to M0 in men with low baseline risk of metastases combined with MI of missing M stage and other variables in all other men. These findings are also relevant for other cancers, if tailored to the context of interest, since the incidence of metastatic cancer is an important proxy for long term cancer-specific mortality in many cancer studies with short follow-up.

Acknowledgments

This project was made possible by the continuous work of the National Prostate Cancer Register of Sweden (NPCR) steering group: David Robinson (register holder) Ingela Franck Lissbrant (chair), Johan Styrke (cochair), Johan Stranne, Jon Kindblom, Camilla Thellenberg, Andreas Josefsson, Ingrida Verbiene, Hampus Nugin, Stefan Carlsson, Anna Kristiansen, Mats Andén, Thomas Jiborn, Olof Ståhl, Olof Akre, Per Fransson, Eva Johansson, Magnus Törnblom, Fredrik Jäderling, Marie Hjälm Eriksson, Lotta Renström, Jonas Hugosson, Ola Bratt, Maria Nyberg, Fredrik Sandin, Camilla Byström, Mia Brus, Mats Lambe, Anna Hedström, Nina Hageman, Christofer Lagerros, Hans Joelsson, and Gert Malmberg.

Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jclinepi.2022.12.008.

References

[1] Gurney J, Sarfati D, Stanley J, Dennett E, Johnson C, Koea J, et al. Unstaged cancer in a population-based registry: prevalence, predictors and patient prognosis. Cancer Epidemiol 2013;37(4):498−504.

[2] Parry MG, Sujenthiran A, Cowling TE, Charman S, Nossiter J, Aggarwal A, et al. Imputation of missing prostate cancer stage in English cancer registry data based on clinical assumptions. Cancer Epidemiol 2019;58:44−51.

[3] Makarov DV, Loeb S, Ulmert D, Drevin L, Lambe M, Stattin P. Prostate cancer imaging trends after a nationwide effort to discourage inappropriate prostate cancer imaging. J Natl Cancer Inst 2013;105:1306−13.

[4] Sobin LH, Compton CC. TNM seventh edition: what's new, what's changed: communication from the International Union against Cancer and the American Joint Committee on Cancer. Cancer 2010;116(22):5336−9.

[5] Sobin LH. In: Sobin LH, Gospodarowicz MK, Wittekind Ch, editors. TNM classification of malignant tumours/. 7th ed..

[6] Mohan K, Pearl J. Graphical models for processing missing data. J Am Stat Assoc 2021;116:1023−37.

[7] Hughes RA, Heron J, Sterne JAC, Tilling K. Accounting for missing data in statistical analyses: multiple imputation is not always the answer. Int J Epidemiol 2019;48:1294−304.

[8] Hardt J, Herke M, Leonhart R. Auxiliary variables in multiple imputation in regression with missing X: a warning against including too many in small sample research. BMC Med Res Methodol 2012;12:184.

[9] RATTEN. Interactive on line report from NPCR. 2022. Available at https://statistik.incanet.se/npcr/. Accessed February 28, 2022.

[10] Van Hemelrijck M, Wigertz A, Sandin F, Garmo H, Hellström K, Fransson P, et al. Cohort profile: the National prostate cancer register of Sweden and prostate cancer data Base Sweden 2.0. Int J Epidemiol 2013;42:956−67.

[11] Statistics Sweden. 2021. Available at https://www.statistikdatabasen.scb.se. Accessed February 28, 2022.

[12] Regionala cancercentrum i samverkan. Available at https://kunskapsbanken.cancercentrum.se/globalassets/cancerdiagnoser/prostatacancer/vardprogram/nvp-prostatacancer.pdf. Accessed February 22, 2022.

[13] Fleiss JL, Levin B, Paik MC. Statistical methods for rates and proportions. Hoboken, NJ: John Wiley & Sons Inc; 2003.

[14] Little RJ, Rubin DB. Statistical analysis with missing data. Hoboken, NJ: John Wiley & Sons; 2019:793.

[15] Buuren S, Groothuis-Oudshoorn C. MICE: multivariate imputation by chained equations in R. J Stat Softw 2011;45:1−67.

[16] White IR, Royston P, Wood AM. Multiple imputation using chained equations: issues and guidance for practice. Stat Med 2011;30:377−99.

[17] Toh S, García Rodríguez LA, Hernán MA. Analyzing partially missing confounder information in comparative effectiveness and safety research of therapeutics. Pharmacoepidemiol Drug Saf 2012;21(Suppl 2):13−20.

[18] Briganti A, Passoni N, Ferrari M, Capitanio U, Suardi N, Gallina A, et al. When to perform bone scan in patients with newly diagnosed prostate cancer: external validation of the currently available guidelines and proposal of a novel risk stratification tool. Eur Urol 2010;57(4):551−8.

[19] Widmark A, Klepp O, Solberg A, Damber JE, Angelsen A, Fransson P, et al. Endocrine treatment, with or without radiotherapy, in locally advanced prostate cancer (SPCG-7/SFUO-3): an open randomised phase III trial. Lancet 2009;373:301−8.

[20] Warde P, Mason M, Ding K, Kirkbride P, Brundage M, Cowan R, et al. Combined androgen deprivation therapy and radiation therapy for locally advanced prostate cancer: a randomised, phase 3 trial. Lancet 2011;378:2104−11.

[21] Sterne JA, White IR, Carlin JB, Spratt M, Royston P, Kenward MG, et al. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. Bmj 2009;338:b2393.

[22] Buuren SV. Flexible imputation of missing data. Boca Raton, FL: Chapman & Hall/CRC Interdisciplinary Statistics; 2018.

[23] Klassen AC, Curriero F, Kulldorff M, Alberg AJ, Platz EA, Neloms ST. Missing stage and grade in Maryland prostate cancer surveillance data, 1992-1997. Am J Prev Med 2006;30(2 Suppl): S77−87.

[24] Merrill RM, Sloan A, Anderson AE, Ryker K. Unstaged cancer in the United States: a population-based study. BMC Cancer 2011;11(1): 402.

[25] Luo Q, Yu XQ, Cooke-Yarborough C, Smith DP, O'Connell DL. Characteristics of cases with unknown stage prostate cancer in a population-based cancer registry. Cancer Epidemiol 2013;37(6): 813−9.

[26] Elliott SP, Johnson DP, Jarosek SL, Konety BR, Adejoro OO, Virnig BA. Bias due to missing SEER data in D'Amico risk stratification of prostate cancer. J Urol 2012;187:2026−31.

[27] Luo Q, Egger S, Yu XQ, Smith DP, O'Connell DL. Validity of using multiple imputation for "unknown" stage at diagnosis in population-based cancer registry data. PLoS One 2017;12:e0180033.

[28] Schröder FH, Hugosson J, Carlsson S, Tammela T, Määttänen L, Auvinen A, et al. Screening for prostate cancer decreases the risk of developing metastatic disease: findings from the European Randomized Study of Screening for Prostate Cancer (ERSPC). Eur Urol 2012;62(5):745−52.

[29] Tomic K, Westerberg M, Robinson D, Garmo H, Stattin P. Proportion and characteristics of men with unknown risk category in the National Prostate Cancer Register of Sweden. Acta Oncologica 2016; 55(12):1461−6.

[30] Tomic K, Sandin F, Wigertz A, Robinson D, Lambe M, Stattin P. Evaluation of data quality in the National prostate cancer register of Sweden. Eur J Cancer 2015;51(1):101−11.

[31] Marshall A, Altman DG, Holder RL. Comparison of imputation methods for handling missing covariate data when fitting a Cox proportional hazards model: a resampling study. BMC Med Res Methodol 2010;10:1−10.

[32] Tateishi U, Morita S, Taguri M, Shizukuishi K, Minamimoto R, Kawaguchi M, et al. A meta-analysis of (18)F-Fluoride positron emission tomography for assessment of metastatic bone tumor. Ann Nucl Med 2010;24(7):523−31.

[33] Palmedo H, Marx C, Ebert A, Kreft B, Ko Y, Türler A, et al. Whole-body SPECT/CT for bone scintigraphy: diagnostic value and effect on patient management in oncological patients. Eur J Nucl Med Mol Imaging 2014;41(1):59−67.

[34] Orrason AW, Westerberg M, Garmo H, Lissbrant IF, Robinson D, Stattin P. Changes in treatment and mortality in men with locally advanced prostate cancer between 2000 and 2016: a nationwide, population-based study in Sweden. BJU Int 2020;126:142−51.

[35] Westerberg M, Franck Lissbrant I, Damber JE, Robinson D, Garmo H, Stattin P. Temporal changes in survival in men with de novo metastatic prostate cancer: nationwide population-based study. Acta Oncologica 2020;59(1):106−11.

[36] Cazzaniga W, Garmo H, Robinson D, Holmberg L, Bill-Axelson A, Stattin P. Mortality after radical prostatectomy in a matched contemporary cohort in Sweden compared to the Scandinavian Prostate Cancer Group 4 (SPCG-4) study. BJU Int 2019;123:421−8.

[37] Epstein JI, Egevad L, Amin MB, Delahunt B, Srigley JR, Humphrey PA. The 2014 international society of urological pathology (ISUP) consensus conference on Gleason grading of prostatic carcinoma: definition of grading patterns and proposal for a new grading system. Am J Surg Pathol 2016;40:244−52.