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

Marcus Westerberga, Kerri Beckmann, Rolf Gedeborg, Sandra Irenaeus, Lars Holmberg, Hans Garmo, Par Stattin

Journal of Clinical Epidemiology 2023

Pokket Sirisreetreerux

Introduction

- The incidence of de novo metastatic cancer (i.e., metastatic cancer at diagnosis) is an early proxy for cancer specific mortality when evaluating intervention.

Introduction

- Missing data on tumor node metastasis (TNM) variables is common 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.

Introduction

- Missing data may also vary over time 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.
- 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.

Aim of the study

- 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.

Materials

- All men diagnosed with prostate cancer from 2000 to 2019 registered in the National Prostate Cancer Register (NPCR) of Sweden were included.
- 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.

Materials

- The following variables were extracted from PCBaSe:
  - age at diagnosis and year of diagnosis,
  - serum level of prostatespecific 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).

Materials

- 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)
  - other or unknown treatment (other)
- Data on all men alive each year between the ages of 40 and 100 years were obtained from Statistics Sweden (SCB).

Materials

- 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 ≥ 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 author estimated the age-standardized incidence of de novo metastatic prostate cancer according to the age distribution in Sweden 2000 by using direct standardization.

Methods

- 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 they used four different methods based on deterministic imputation (DI) and multiple imputation (MI) using the R package mice.
- The number of MIs was set to 128.
- M stage was considered missing if the man had not undergone imaging to assess metastatic status.

Methods

- 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.

Methods

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.

Methods

2.Partial deterministic imputation + multiple imputation

- For men with low-risk prostate cancer 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.

Methods

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.

Methods

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.

Results

1. Baseline characteristics (Table 1)
    - There were 190,420 men diagnosed with prostate cancer between 2000 and 2019 in NPCR.
    - Of which 126,102 men (66%) had missing M stage; 15,526 men (8%) were M1, constituting 24% of all imaged men.

Results

1. Baseline characteristics (Table 1)

- 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.

	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								
2000–2005	50,744	(27)	14,894	(31) [29]	4,955	(32) [10]	30,895	(25) [61]
2006–2011	56,868	(30)	10,005	(21) [18]	3,527	(23) [6]	43,336	(34) [76]
2012–2019	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]

Results

1. Baseline characteristics
   - Thirty six percent of men with M0 and 3% of men with M1 were categorized as low metastatic risk.
   - 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.

	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								
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]

Results

2. Baseline characteristics after imputation
   - The proportions of men with imputed M1 among men with missing M stage were
     - 7%, PDI+MI
     - 10%, standard MI (SMI)
     - 16% restricted MI (RMI)

Results

2. Baseline characteristics after imputation

- Among men with imputed M1, the proportion categorized as low metastatic risk varied substantially (1-40%) depending on the imputation method used, compared with 4% among men with known M1.

		M stage M0						M stage M1					
		Partial deterministic imputation + MI		Standard MI		Restricted MI		Partial deterministic imputation + MI		Standard MI		Restricted MI	
		n	(%)	n	(%)	n	(%)	n	(%)	n	(%)	n	(%)
Metastatic risk*													
Low metastatic risk	93059	(79)	90742	(79)	82003	(77)	90	(1)	2380	(20)	7711	(40)	
High metastatic risk	24222	(21)	23650	(21)	20219	(19)	8730	(99)	9330	(80)	9954	(52)	

Results

2. Baseline characteristics after imputation

- 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.
  - Different some baseline characteristics with different methods

		M stage M1					
		Partial deterministic imputation + MI		Standard MI		Restricted MI	
		n	(%)	n	(%)	n	(%)
N		8820	(100)	11710	(100)	19279	(100)
Age at diagnosis, years							
<60		195	(2)	472	(4)	1188	(6)
60-69		1054	(12)	2022	(17)	4054	(21)
70-74		1211	(14)	1744	(15)	2908	(15)
75-80		1893	(21)	2408	(21)	3689	(19)
80+		4467	(51)	5065	(43)	7441	(39)
Year of diagnosis							
2000-2005		3360	(38)	3958	(34)	6447	(33)
2006-2011		3227	(37)	3975	(34)	6774	(35)
2012-2019		2233	(25)	3777	(32)	6059	(31)
Charleson Comorbidity Index							
0		4629	(52)	6479	(55)	12083	(63)
1		1766	(20)	2203	(19)	3218	(17)
2		1285	(15)	1622	(14)	2264	(12)
3+		1140	(13)	1406	(12)	1714	(9)
PSA (ng/mL)							
0-9		445	(5)	2062	(18)	7333	(38)
10-19		796	(9)	1656	(14)	3776	(20)
20-49		1968	(22)	2320	(20)	3085	(16)
50-99		1663	(19)	1755	(15)	1655	(9)
100-499		2609	(30)	2599	(22)	1850	(10)
500+		1339	(15)	1318	(11)	800	(4)
Missing		0	(0)	0	(0)	781	(4)
T stage							
1		850	(10)	2590	(22)	5337	(28)
2		2047	(23)	3022	(26)	5503	(29)
3		4357	(49)	4517	(39)	6522	(34)
4		1566	(18)	1581	(13)	1917	(10)
Missing		0	(0)	0	(0)	0	(0)

Results

2. Baseline characteristics after imputation

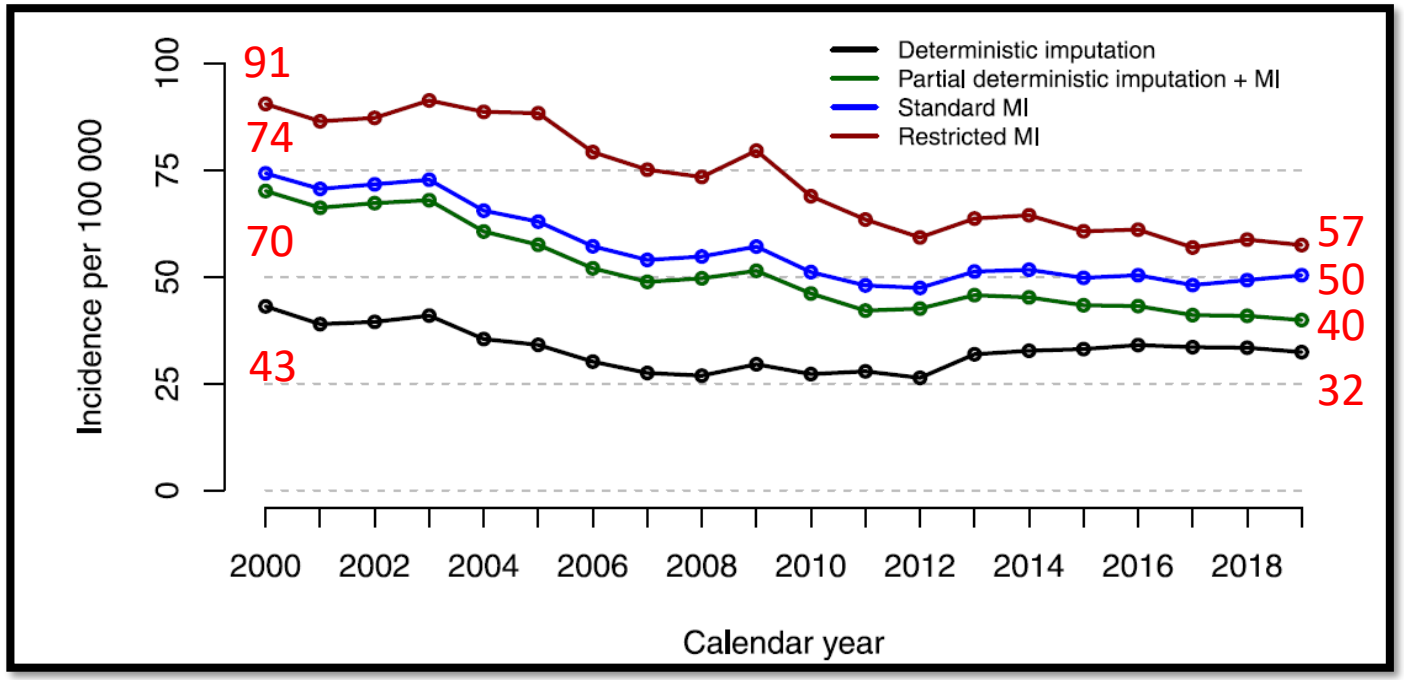
- The tumor characteristics among men with imputed M0 were similar across methods and tended toward more favorable disease characteristics compared to men with known M0.

	M stage M0					
	Partial deterministic imputation + MI		Standard MI		Restricted MI	
	n	(%)	n	(%)	n	(%)
N	117282	(100)	114392	(100)	106823	(100)
Age at diagnosis, years						
<60	17300	(15)	17023	(15)	16307	(15)
60-69	47454	(40)	46486	(41)	44454	(42)
70-74	21268	(18)	20735	(18)	19571	(18)
75-80	15939	(14)	15424	(13)	14143	(13)
80+	15321	(13)	14723	(13)	12347	(12)
Year of diagnosis						
2000-2005	27535	(23)	26937	(24)	24448	(23)
2006-2011	40109	(34)	39361	(34)	36562	(34)
2012-2019	49638	(42)	48094	(42)	45812	(43)
Charleson Comorbidity Index						
0	87164	(74)	85314	(75)	79710	(75)
1	14187	(12)	13750	(12)	12735	(12)
2	9866	(8)	9529	(8)	8887	(8)
3+	6065	(5)	5799	(5)	5491	(5)
PSA (ng/mL)						
0-9	78871	(67)	77222	(68)	69812	(65)
10-19	23806	(20)	22945	(20)	20064	(19)
20-49	9629	(8)	9296	(8)	7926	(7)
50-99	3050	(3)	2974	(3)	2849	(3)
100-499	1709	(1)	1718	(2)	2280	(2)
500+	217	(0)	237	(0)	731	(1)
Missing	0	(0)	0	(0)	3160	(3)
T stage						
1	72523	(62)	70781	(62)	67949	(64)
2	34099	(29)	33128	(29)	30699	(29)
3	9846	(8)	9687	(8)	7728	(7)
4	813	(1)	796	(1)	447	(0)
Missing	0	(0)	0	(0)	0	(0)
Gleason score						
Gleason score 6	65813	(56)	63988	(56)	57713	(54)
Gleason score 7	41341	(35)	40311	(35)	35602	(33)
Gleason score 8-10	10128	(9)	10093	(9)	9202	(9)
Missing	0	(0)	0	(0)	4306	(4)
N stage						
0	109512	(93)	106888	(93)	99982	(94)
1	7770	(7)	7504	(7)	6841	(6)
Missing	0	(0)	0	(0)	0	(0)

Results

3. Incidence of metastatic prostate cancer

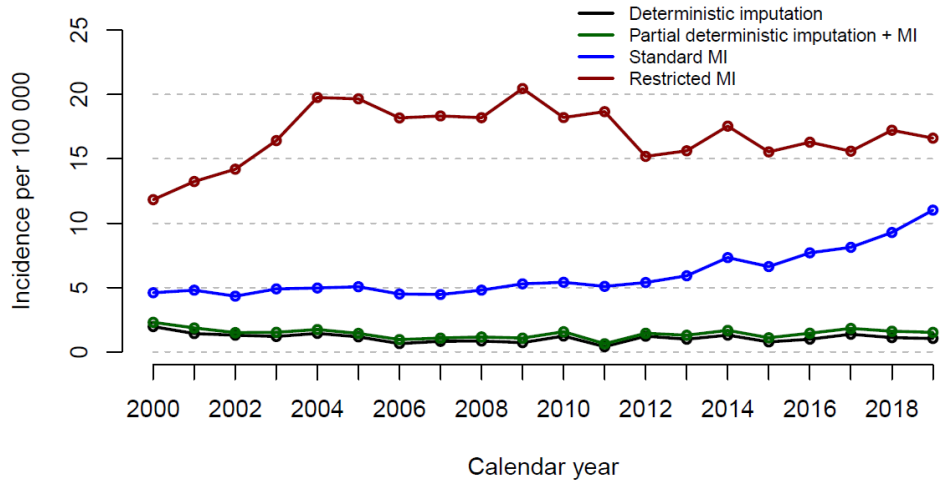
- The estimated age-standardized incidence of de novo metastatic prostate cancer varied markedly between the four applied methods.



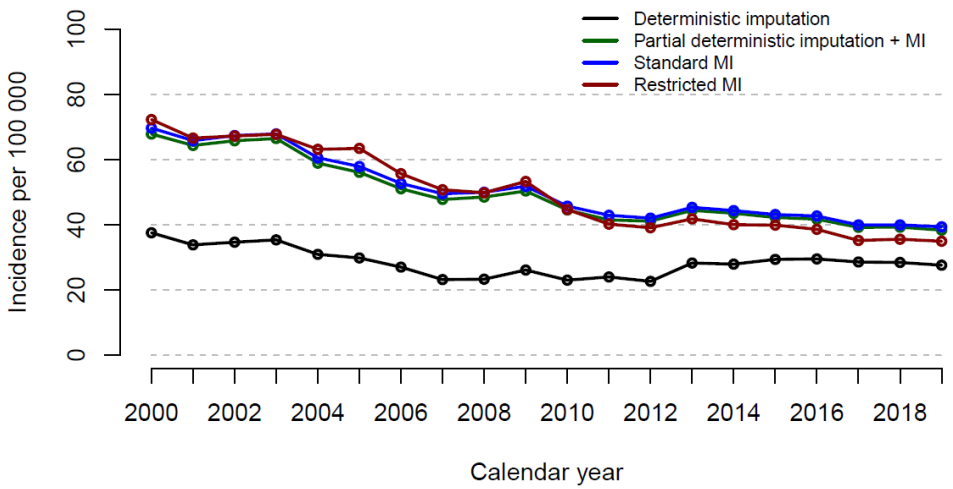
Both the estimated incidences, as well as the difference in estimated incidences between methods, decreased with time.

Results

3. Incidence of metastatic prostate cancer
   - The estimated annual incidence of men with de novo metastatic prostate cancer categorized as low metastatic risk varied between methods.
   - The estimated annual incidence of men with de novo metastatic prostate cancer categorized as high metastatic risk was similar for all methods except DI.



Low metastatic risk



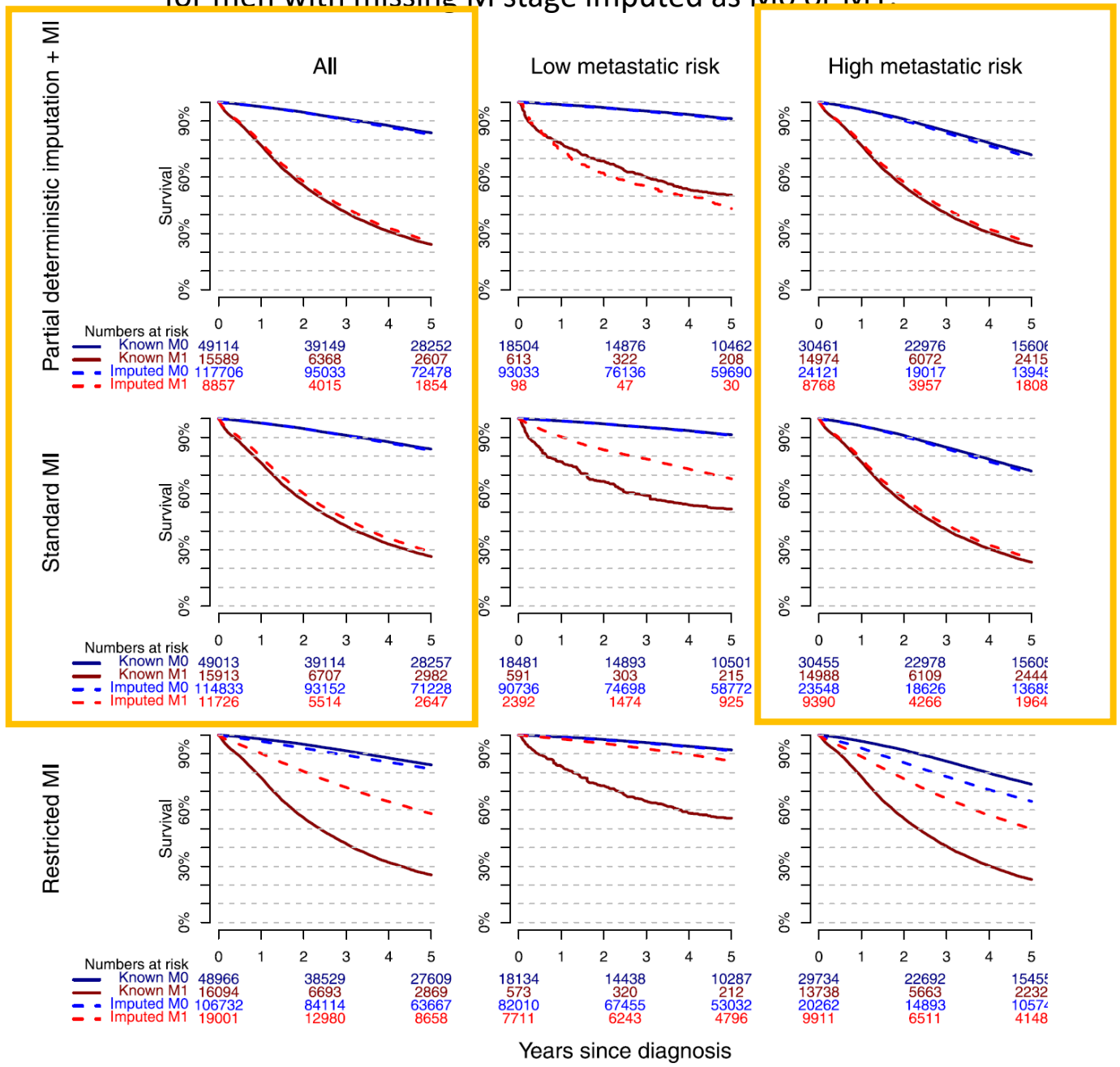
High metastatic risk

Results

4. Survival

- 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.

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.

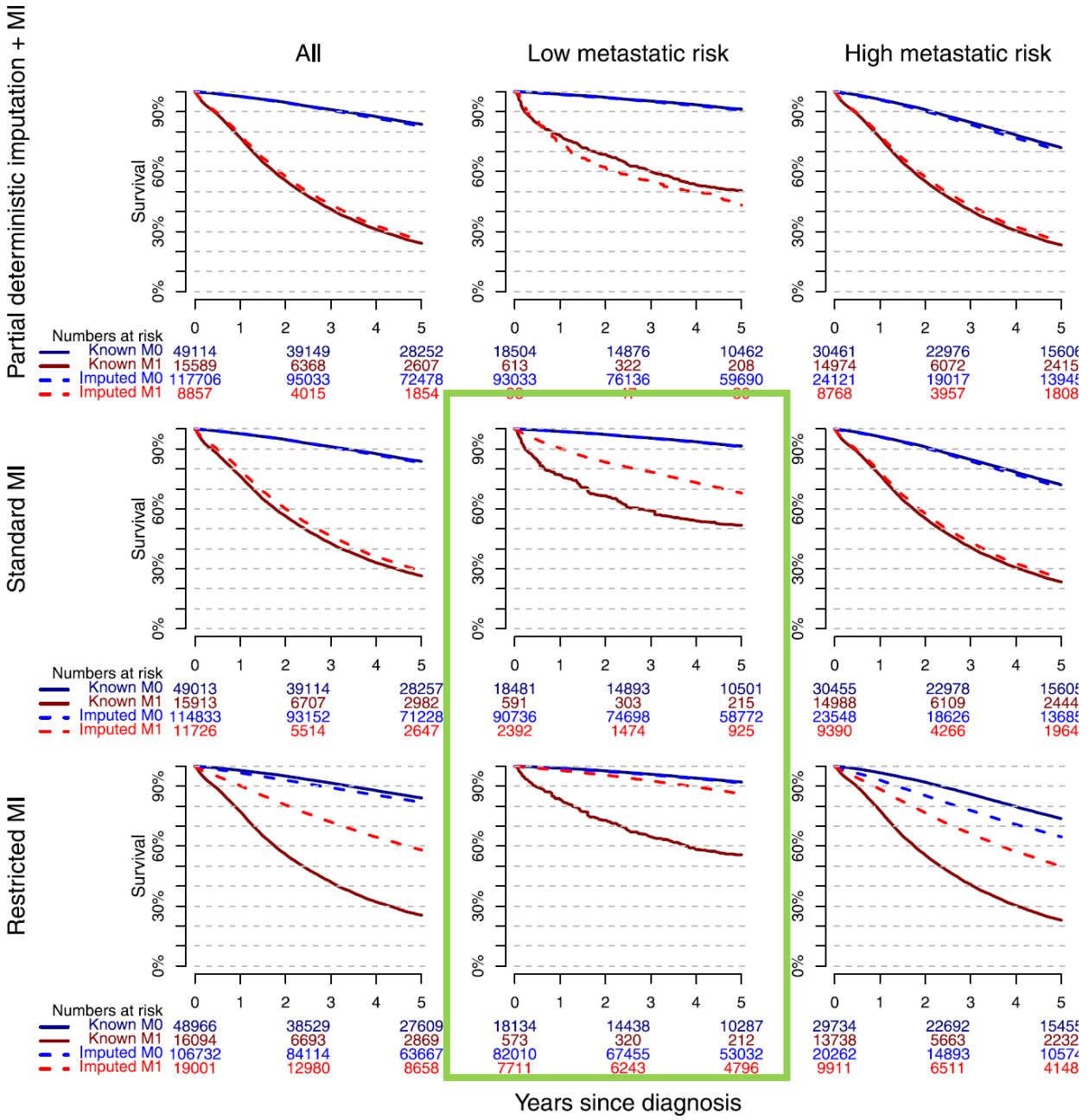


Results

4. Survival

- 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.
- For men with known and imputed M1 categorized as low metastatic risk, the curves 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.

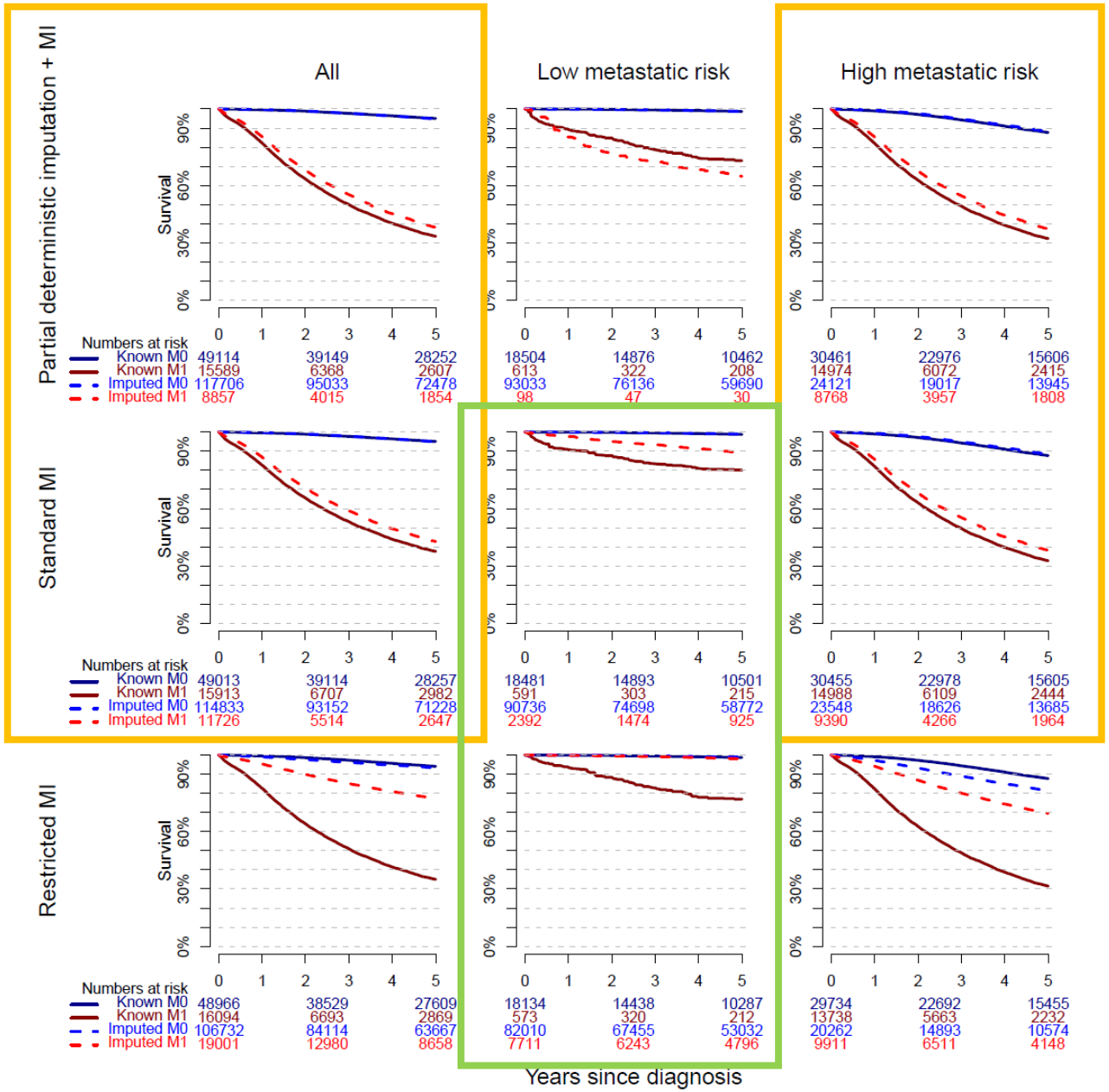
The adjusted 5-year overall survival curves for men with known M0 or M1, and for men with missing M stage imputed as M0 or M1.



Results

- The results were similar for prostate cancer specific survival.

Adjusted prostate cancer-specific survival averaged over the multiple imputations, for all men and stratified by those with Low metastatic risk and High metastatic risk.



Discussion

- Summary of findings
  - The estimated age-standardized incidence of de novo metastatic prostate cancer differed markedly between the methods used to handle missing data in metastatic status.
  - PDI+MI 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.

Discussion

- 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.

  - The validity of the MI methods relies on the plausibility of the missing at random (MAR) assumption.
  - It is recommended to include as many auxiliary variables as possible in the analysis to increase the plausibility of MAR, since such variables may explain systematic differences between those with observed and missing data.

Discussion

- When such variables are not available or omitted, data can no longer be considered MAR and is instead missing not at random (MNAR).
- 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.

Discussion

- 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.

Discussion

- 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.

Discussion

- Strengths
  - Data quality in NPCR has been shown to be high.
  - 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.

Discussion

- Limitations
  - The large proportion of missing data in M stage (66%) and missing data are predictors for imputing M stage that may affect the performance of MI.
  - The author were unable to assess the potential bias of different use of imagings modalities, due to lack of such data.
  - 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.

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.

Thank you