Cancer Epidemiology 58 (2019) 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]

[a] Department of Health Services Research and Policy, London School of Hygiene and Tropical Medicine, 15-17 Tavistock Place, London, WC1H 9SH, England, United Kingdom
[b] Clinical Effectiveness Unit, The Royal College of Surgeons of England, 35-43 Lincoln's Inn Fields, London, WC2A 3PE, England, United Kingdom
[c] Department of Radiotherapy, Guy's and St Thomas' NHS Foundation Trust, Great Maze Pond, London, SE1 9RT, England, United Kingdom
[d] Department of Cancer Epidemiology, Population, and Global Health, King's College London, Strand, London, WC2R 2LS, England, United Kingdom
[e] Department of Urology, The Christie NHS Foundation Trust, Wilmslow Road, Manchester, M20 4BX, England, United Kingdom
[f] Department of Urology, Salford Royal NHS Foundation Trust, Stott Lane, Salford, M6 8HD, England, United Kingdom
[g] Department of Oncology, University College London Hospitals, 235 Euston Road, London, NW1 2BU, England, United Kingdom

ARTICLE INFO

Keywords:
Prostate cancer
Neoplasm staging
Risk stratification
Clinical imputation
Missing data
Unknown stage
Epidemiology
Cancer registry
Population-based

ABSTRACT

Background: Cancer stage can be missing in national cancer registry records. We explored whether missing prostate cancer stage can be imputed using specific clinical assumptions.

Methods: Prostate cancer patients diagnosed between 2010 and 2013 were identified in English cancer registry data and linked to administrative hospital and mortality data (n = 139,807). Missing staging items were imputed based on specific assumptions: men with recorded N-stage but missing M-stage have no distant metastases (M0); low/intermediate-risk men with missing N- and/or M-stage have no nodal disease (N0) or metastases; and high-risk men with missing M-stage have no metastases. We tested these clinical assumptions by comparing 4-year survival in men with the same recorded and imputed cancer stage. Multi-variable Cox regression was used to test the validity of the clinical assumptions and multiple imputation.

Results: Survival was similar for men with recorded N-stage but missing M-stage and corresponding men with M0 (89.5% vs 89.6%); for low/intermediate-risk men with missing M-stage and corresponding men with M0 (92.0% vs 93.1%); and for low/intermediate-risk men with missing N-stage and corresponding men with N0 (90.9% vs 93.7%). However, survival was different for high-risk men with missing M-stage and corresponding men with M0. Imputation based on clinical imputation performs as well as statistical multiple imputation.

Conclusion: Specific clinical assumptions can be used to impute missing information on nodal involvement and distant metastases in some patients with prostate cancer.

1. Introduction

As prostate cancer is biologically heterogeneous, it is important to differentiate clinically between indolent, low-risk disease that is localised to the prostate and disease that is highly aggressive and likely to metastasise [1]. Management options for non-metastatic disease differ significantly depending on the clinical stage and grade of disease, ranging from active surveillance to radical local treatment (with or without supplementary systemic therapy). Patients presenting with distant metastases are managed differently with systemic and palliative treatments taking precedence [2].

Cancer registry and administrative hospital data provide a vast source of cancer-specific information [3]. These data are used extensively in population-based studies to investigate nationwide trends in cancer diagnosis and management [3–5]. When prostate cancer-specific information is not known, or simply not recorded, a patient's cancer stage often cannot be determined. This is potentially problematic when these data are being used for the purposes of research and

* Corresponding author at: Clinical Effectiveness Unit, Royal College of Surgeons of England, 35-43 Lincoln's Inn Fields, London, WC2A 3PE, England, United Kingdom.
E-mail addresses: mparry@rcseng.ac.uk (M.G. Parry), ASujenthiran@rcseng.ac.uk (A. Sujenthiran), Thomas.Cowling@lshtm.ac.uk (T.E. Cowling), susan.charman@cysticfibrosis.org.uk (S. Charman), jnossiter@rcseng.ac.uk (J. Nossiter), Ajay.Aggarwal@lshtm.ac.uk (A. Aggarwal), Noel.Clarke@srft.nhs.uk (N.W. Clarke), heather_payne@blueyonder.co.uk (H. Payne), Jan.vanderMeulen@lshtm.ac.uk (J. van der Meulen).

https://doi.org/10.1016/j.canep.2018.11.003
Received 1 August 2018; Received in revised form 5 November 2018; Accepted 9 November 2018
Available online 18 November 2018
1877-7821/

health service evaluation, especially when staging greatly affects prognosis and clinical management [3,6,7].

In many countries completeness of cancer registry data is improving over time but there is still a reliance on historical data, which are important in studying long-term outcomes and survival trends in prostate cancer [8,9]. The necessary omission of patients with incomplete data from analyses may introduce significant selection bias and will lead to loss of statistical power.

Statistical methods, such as multiple imputation, are often used when dealing with missing data. These techniques re-classify missing values based on the patterns present in other variables within the sample. [4,10–14]. Staging data is often missing in certain clinical situations. For example, in prostate cancer registration, negative results for nodal involvement (N0) or distant metastases (M0) may be more likely to be missing than positive results (N1 and M1, respectively). This is a consequence of two factors: first, the staging practice, as patients with low/intermediate-risk prostate cancer may not have further staging investigations and second, record keeping, as negative results are less likely to be recorded than positive results.

Given this pattern of missingness, imputation based on specific 'clinical assumptions' rather than statistical imputation provides an alternative approach to improve the completeness of prostate cancer staging information. For example, prostate cancer patients with tumour stage T1 and a Gleason score of six are unlikely to undergo staging investigations because the likelihood they have nodal involvement or distant metastases is very low. For these patients, one can assume that the missing N- and M-stage are N0 and M0, respectively. Similar assumptions can be made for patients who have complete data for nodal status but missing data for distant metastases. If staging data is available for nodal disease, it is likely that staging was also performed to look for distant metastases but the negative result was not recorded. The likelihood of missing M-stage representing positive disease is therefore very low and missing M-stage can be assumed to be M0.

Specific clinical assumptions may therefore be used to impute missing staging items in cancer registry data in a systematic way in order to risk stratify more men with prostate cancer. We explored to what extent staging data completeness can be improved by using the following four clinical assumptions:

1 Men with a recorded N-stage but missing M-stage have no distant metastases (M0).
2 Low/intermediate-risk men with missing M-stage have no distant metastases (M0).
3 Low/intermediate-risk men with missing N-stage have no nodal disease (N0).
4 High-risk men with missing M-stage have no distant metastases (M0).

An assumption for high-risk men and missing N-stage was not assessed given that for the majority of research and evaluation purposes these men will be already assigned to the locally advanced group as a minimum.

2. Material and methods

All patients who were diagnosed with prostate cancer between January 1st 2010 and December 31st 2013 were identified in the English cancer registry using the ICD-10 code 'C61' [15]. Data collected by the eight regional English cancer registries have been combined into a national data set. This was then linked at patient-level to Hospital Episode Statistics (HES), an administrative database of all hospital admissions in the English National Health Service and data from the Office for National Statistics (ONS), giving 139,807 patients over this time period. Follow-up was available to 31st December 2014. Data collected from the English cancer registry included age, Gleason score, and T-, N- and M-stages (TNM). TNM data used preferentially clinical cancer registry items and then pathological cancer registry items, in line with the Union for International Cancer Control (UICC) TNM 7th edition, taking staging information that was updated as much as possible by cancer registry staff.

Table 1
Staging criteria used for prostate cancer according to the National Prostate Cancer Audit without clinical imputation.

Prostate Cancer Stage	Staging Criteria
Advanced	M1
Locally Advanced	(≥T3 or GS ≥8 or N1) + M0
Localised	≤T2 + GS ≤7 + N0 + M0
Unknown Stage	None of the above

The Royal College of Surgeons (RCS) Charlson score was used to identify co-morbid conditions from the HES records based on previously coded co-morbidity within one year of their prostate cancer diagnosis [16]. Socio-economic deprivation was derived from the Index of Multiple Deprivation (IMD, The English Indices of Deprivation, 2012). The IMD ranks 32,482 areas in the country and patients are grouped into socioeconomic quintiles based on the national distribution.

Men were labelled as 'complete' if their prostate cancer could be clinically categorised into localised, locally advanced or advanced disease (Table 1). Low/intermediate-risk, non-metastatic prostate cancer was defined as T1 or T2 and Gleason score ≤7 in the absence of nodal or distant metastases. High-risk prostate cancer was defined as any one of T3, T4, N1 or Gleason score ≥8 in the absence of distant metastases (M0). According to the National Prostate Cancer Audit [17] and the National Institute for Health and Care Excellence [2] risk stratification, high-risk localised disease was classified as 'locally advanced' disease. PSA was not available in the English cancer registry prior to 2014 and so was not used. Patients with missing data items were only considered to have a missing stage if there was insufficient data elsewhere to clinically stage their disease. For example, a patient with documented metastases (M1) can be staged as advanced, irrespective of the completeness of the other stratification variables. Staging completeness was stratified by year of diagnosis, age, RCS Charlson score and socio-economic deprivation.

Kaplan Meier analysis was used to compare 4-year overall survival in men with missing and complete staging items according to each of our clinical assumptions (see above). Patient survival was displayed with corresponding 95% confidence intervals to compare survival in patients with complete and imputed staging data. The log-rank test was used to test differences between survival curves at a p-value of 0.05. To be consistent with each clinical assumption we expect that:

1 Men with missing M-stage or M0 will have similar survival if the patient has a recorded N-stage.
2 Low/intermediate-risk men with missing M-stage or M0 will have similar survival.
3 Low/intermediate-risk men with missing N-stage or N0 will have similar survival.
4 High-risk men with missing M-stage or M0 will have similar survival.

'X' values in the English cancer registry represent when N- and M-stage information was either inconclusive or missing (NX and MX, respectively). In this analysis, we used NX and MX to represent missing values.

Multi-variable Cox regression was used to explore survival time according to each of the clinical assumptions, adjusting for age, RCS Charlson score, socio-economic deprivation, year of diagnosis and the eight English cancer registry regions displaying hazard ratios (HRs) and 95% confidence intervals (CIs). We then compared this clinical imputation method with that of multiple imputation by chained

Table 2
Completeness of prostate cancer staging with and without clinical imputation.

			Staging completeness						
			Without assumptions				With assumptions		
	N	%	Metastatic n (%)	Locally advanced n (%)	Localised n (%)	Missing n (%)	Locally advanced n (%)	Localised n (%)	Missing n (%)
Total	139,807	100	13,257 (9.5)	24,889 (17.8)	22,341 (16.0)	79,320 (56.7)	27,899 (20.0)	39,960 (28.6)	58,691 (42.0)
Year of diagnosis									
2010	33,701	24.1	2009 (6.0)	2845 (8.4)	841 (2.5)	28,006 (83.1)	3589 (10.7)	4544 (13.5)	23,559 (69.9)
2011	33,463	23.9	2361 (7.1)	4354 (13.0)	2089 (6.2)	24,659 (73.7)	5283 (15.8)	6355 (19.0)	19,464 (58.2)
2012	35,225	25.2	4370 (12.4)	8130 (23.1)	8239 (24.4)	14,486 (41.1)	8741 (24.8)	12,677 (36.0)	9437 (26.8)
2013	37,418	26.8	4517 (12.1)	9560 (25.6)	11,172 (29.9)	12,169 (32.5)	10,286 (27.5)	16,384 (43.8)	6231 (16.7)
Age group (yrs)									
≤60	15,321	11.0	856 (5.6)	2375 (15.5)	3571 (23.3)	8519 (55.6)	2822 (18.4)	6151 (40.2)	5492 (35.9)
60–69	47,923	34.3	3136 (6.5)	8972 (18.7)	9661 (20.2)	26,154 (54.6)	10,341 (21.6)	16,783 (35.0)	17,663 (36.9)
70–79	51,063	36.5	4740 (9.3)	10,102 (19.8)	7904 (15.5)	28,317 (55.5)	11,093 (21.7)	14,049 (27.5)	21,181 (41.5)
80–89	22,435	16.1	3819 (17.0)	3.184 (14.2)	1187 (5.3)	14,245 (63.5)	3371 (15.0)	2907 (13.0)	12,338 (55.0)
≥90	3065	2.2	706 (23.0)	256 (8.4)	18 (0.6)	2085 (68.0)	272 (8.9)	70 (2.3)	2017 (65.8)
Number of co-morbidities (RCS Charlson score)									
0	109,470	78.3	8256 (7.5)	20,469 (18.7)	18,530 (16.9)	62,215 (56.8)	23,011 (21.0)	33,062 (30.2)	45,141 (41.2)
1	21,007	15.0	2961 (14.1)	3285 (15.6)	2987 (14.2)	11,774 (56.1)	3642 (17.3)	5285 (25.2)	9119 (43.4)
> 2	9330	6.7	2040 (21.9)	1135 (12.2)	824 (8.8)	5331 (57.1)	1246 (13.4)	1613 (17.3)	4431 (47.5)
Socio-economic deprivation									
1 (least deprived)	34,259	24.5	2863 (8.4)	5683 (16.6)	5560 (16.2)	20,153 (58.8)	6546 (19.1)	10,532 (30.7)	14,318 (41.8)
2	33,669	24.1	3082 (9.2)	6145 (18.3)	5410 (16.1)	19,032 (56.5)	6860 (20.4)	9961 (29.6)	13,766 (40.9)
3	29,533	21.1	2945 (10.0)	5426 (18.4)	4578 (15.5)	16,584 (56.2)	6060 (20.5)	8311 (28.1)	12,217 (41.4)
4	23,462	16.8	2357 (10.1)	4184 (17.8)	3735 (15.9)	13,186 (56.2)	4625 (19.7)	6354 (27.1)	10,126 (43.2)
5 (most deprived)	18,884	13.5	2010 (10.6)	3451 (18.3)	3058 (16.2)	10,365 (54.9)	3808 (20.2)	4802 (25.4)	8264 (43.8)

equations. Ten complete data sets were created taking account of age, RCS Charlson score, socio-economic deprivation, diagnosis year, T-stage and Gleason score [18]. For the groups of patients defined by each clinical assumption we determined how often missing N and M stages were assigned to be N0 or N1 and M0 or M1 by averaging the proportions of patients with these specific imputed staging results over the ten data sets. We also used Cox regression analysis to calculate separately specific hazard ratios for patients for whom the statistical imputation had assigned N0 or N1 for patients with missing N stage and M0 or M1 for patients with missing M stage. The hazard ratios for these specific imputed results were combined using Rubin's rules.

3. Results

A total of 139,807 men with a diagnosis of prostate cancer in the English cancer registry could be linked to the HES database (January 1st 2010–December 31st 2013) and 43% of the patients could be staged accurately. Completeness and variation of clinical staging is shown in Table 2. Completeness increased with year of diagnosis, from 17% in 2010 to 68% in 2013. Completeness was similar for all age groups below 80 years (43%–45%) and decreased for those aged between 80–90 years (37%) and > 90 years (32%). As socio-economic deprivation increased, data completeness improved slightly from 41% to 45% but the presence of co-morbidity did not appear to affect completeness.

When we explored the validity of imputations of missing N-stage or M-stage based on our four clinical assumptions we found the following results. First, 4-year overall survival was very similar for men with a recorded N-stage, but missing M-stage, and corresponding men with recorded M0 (89.5% vs 89.6%). This similarity was observed to such a degree that the patient survival curves for MX and M0 appear as one line. Survival of men with M1 in this cohort was substantially lower at 39.7% (Fig. 1). This pattern was also observed when the analysis was restricted to N0 and N1 disease individually.

Second, for low/intermediate-risk men with missing M-stage and corresponding men with recorded M0, survival was also very similar (92.0% vs 93.1%), and substantially higher than men with recorded M1 (59.2%). Again, this similarity was observed to such a degree that the patient survival curves for MX and M0 appear as one line (Fig. 2a).

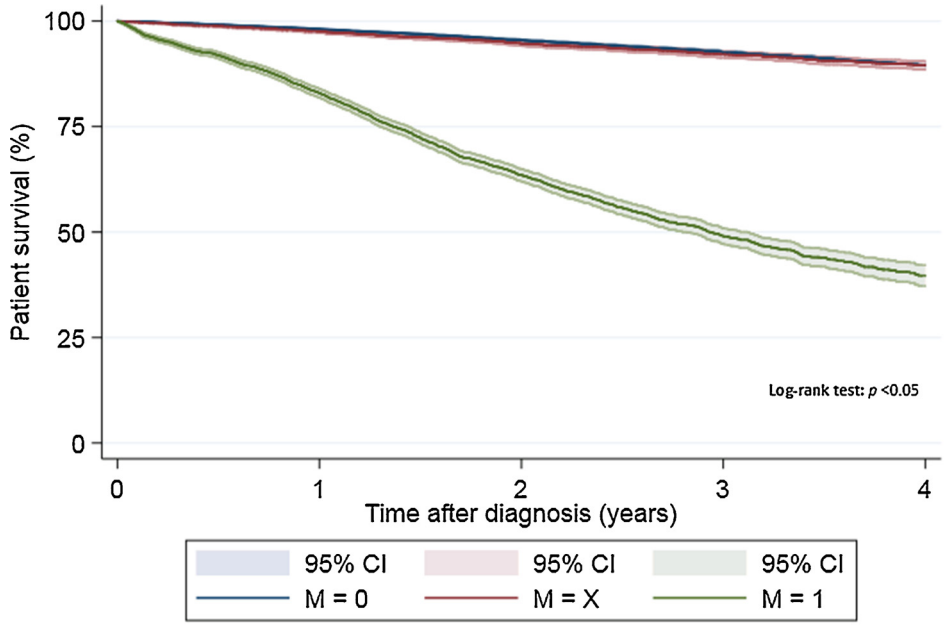
Third, the same pattern was seen for low/intermediate-risk men with missing N-stage and corresponding men with recorded N0 (90.9% vs 93.7%), where survival was also much lower for men with N1 (81.4%, Fig. 2b).

Fourth, 4-year survival for high-risk men with missing M-stage was 71.4% which differed substantially from corresponding men with recorded M0 or M1 (84.5% vs 36.7%, Fig. 3). These results support the first three clinical assumptions that:

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

Assumptions two and three therefore assume that any man with both missing N-stage and missing M-stage are N0 and M0. The fourth clinical assumption was not supported, thus missing M-stage cannot be imputed to M0 in high-risk men. All the 95% confidence intervals presented with the survival curves were narrow around the survival estimates, highlighting the precision of our results and appropriateness of our assumptions. When we used these three clinical assumptions to impute missing N-stage and M-stage, we could increase staging completeness from 43% to 58% by allocating an extra 20,629 patients to an appropriate stage over the four year time frame (2010–2013). For the most recent data only (2013), completeness rose from 68% (without assumptions) to 83% (with assumptions).

Adjusted Cox regression, compared to the unadjusted analysis, showed that there was some residual bias when using the first three clinical assumptions as there were significant differences in survival between MX (assumption one and three) or NX (assumption two) compared to either N0 or M0, respectively (HRs and 95% CIs > 1) (Table 3). The adjusted model provided further support for assumption two as it showed a decrease from an unadjusted HR of 1.50 to an adjusted HR of 1.26, confirming the presence of confounding from other



***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.

Fig. 1. Overall survival for men with complete N-stage (N1/N0) showing the distribution of M-stage (M1/M0/missing M).

variables.

However, the performance of our clinical assumptions compared to multiple imputation is very similar. For assumptions two and three, survival of NX and MX values imputed to N0 and M0 using multiple imputation were in line with our clinical assumptions given the very similar hazard ratios (Tables 3 and 4). Multiple imputation rarely imputes values to M1 (assumption two) or N1 (assumption three), highlighting that our clinical assumptions are as appropriate as multiple imputation in this setting and both methods perform relatively well for patients in the low/intermediate-risk group. For assumption one, our clinical assumptions were weaker than the method using multiple imputation given the hazard ratios for the MX values imputed to M0 were further away from 0 (HR 1.39 95% 1.27–1.53 vs. HR 1.13 95% CI 1.00–1.28). However multiple imputation was still imperfect as those missing values imputed to M1, rather than M0, were not comparable to the survival of men with M1 (HR 2.53 95% CI 2.07–3.09 vs. 7.11 95% CI 6.71–7.53), a bias which does not affect our clinical assumptions given they are assumed to be negative (M0).

4. Discussion

4.1. Overview

Imputation based on specific clinical assumptions increased the completeness of clinical staging from 43% to 58% for prostate cancer patients recorded in the English cancer registry between 2010 and 2013. We found that overall survival of patients with available N- or M-stage were similar to those with imputed results based on three of the four clinical assumptions, thus providing evidence for their validity. These clinical assumptions perform as well as multiple imputation and are more easily applicable at local hospitals for those without appropriate statistical software or expertise.

4.2. Other national databases

Staging completeness is variable between national data sets in Europe, Australasia and the US [8,9,19–24]. The National Prostate Cancer Registry of Sweden reports staging completeness at 97% when using the National Comprehensive Cancer Network (NCCN) risk classification. This represents the highest level of any national data set and uses Gleason score and PSA as well as TNM—a success which is likely to be attributed to cancer registration being compulsory and mandated in law [25]. A further European cancer registry in Denmark reports staging completeness at 71% (2004–2009) restricting classification to TNM [9]. The Australasian databases have published completeness at 58% and 27% in New South Wales, Australia (1999–2007) [8] and New Zealand (2006–2008) [19], respectively, according to the SEER summary stage.

Completeness of tumour variables in the US SEER database, using Medicare patients, is reported at 67% (2004–2006) and is unique by using a detailed D'Amico risk stratification which utilises all of PSA, Gleason and a sub-classification of T2 [22]. T2(a–c) describes the proportion of the prostate affected and provides an additional variable for quantifying risk. Whatever risk stratification is being used, this overview of completeness in national databases indicates that in most countries approaches that are similar to the one that we present in this paper may be used to increase staging completeness of cancer registry data.

4.3. Alternative approaches for dealing with missing staging data

Missing N- and M-stage data is usually dealt with by reclassifying them to N0 and M0. The National Prostate Cancer Registry of Sweden uses a modified version of the NCCN classification and in doing so assumes that men labelled 'not N1' or 'not M1' are N0 and M0, respectively. As we have shown, we could not assume that high-risk men are free from nodal and/or distant metastases. With only 3% of men being unstaged in the Swedish cancer registry, it is unlikely that this approach will have led to substantial biased staging profiles but it may be less appropriate for use in countries that have cancer registries with less complete staging data. In addition, the modified NCCN classification uses PSA to classify regional and distant metastases (≥ 50 and > 100,

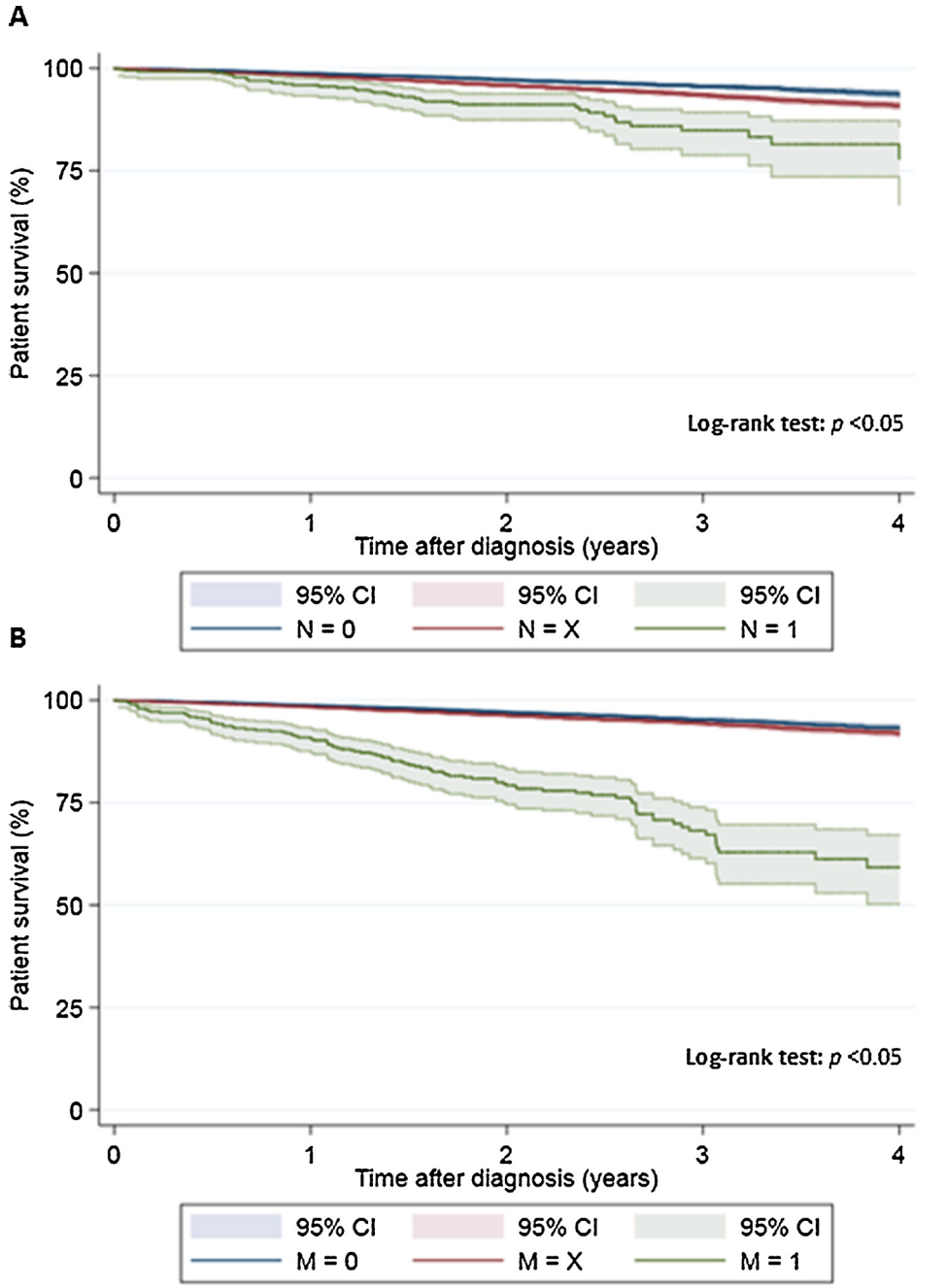


Fig. 2. Overall survival for men with low/intermediate-risk disease (T1-2 and Gleason score ≤7) showing the distribution of: a. M-stage (M1/M0/missing M) b. N-stage (N1/N0/missing N).

respectively) which may prevent the understaging of missing N- and M-stage in Swedish men [24].

A study of the completeness of T, N and M in the Danish cancer registry used an approach which has similarities with the one that we present in this study. The Danish approach led to the staging of 70.5% by assuming that men with tumour stage T1 or 2 and missing N- and M-stage had localised disease, which implies that they assumed these men had N-stage N0 and M-stage M0 [9]. This is very similar to our approach but we were able to include Gleason score in our staging method and validate our assumptions by comparing survival estimates in men with complete and imputed missing staging data.

An Australian population-based cohort study, linked with cancer registry data, concluded that multiple imputation can provide valid estimates for missing staging data but they indicated that caution should be used when applying their methods to other data sets [13]. An important element of their imputation method was that it also used information on the primary treatment that patients received in the first six months after diagnosis, although even without treatment

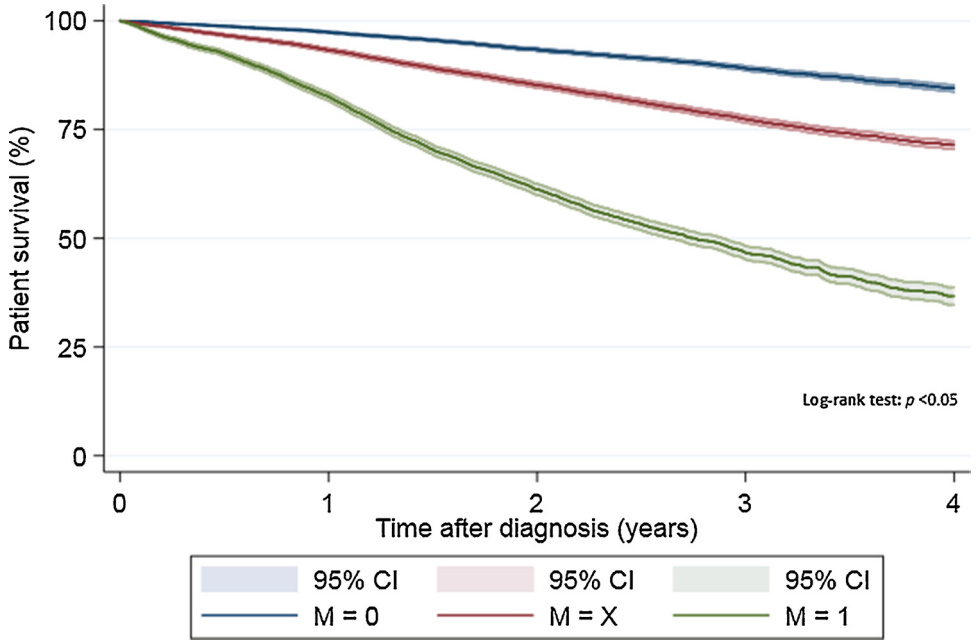


Fig. 3. Overall survival for men with high-risk disease (T3-4 or Gleason score ≥8) showing the distribution of M-stage (M1/M0/missing M).

information included, multiple imputation appears to provide valid estimates.

4.4. Characteristics of incomplete data

Characteristics of men with incomplete prostate cancer staging information have not been previously described using the English cancer registry. Consistent with previous literature, completeness of clinical staging of prostate cancer decreases with increasing age. Other studies have also shown that co-morbidity and socio-economic deprivation is associated with missing data but our data does not support this [9,19–21,26–28]. For the very elderly, the likelihood of undergoing staging investigations is low given their diminished life expectancy. It has also been shown that cancer staging is more often incomplete in patients with cancer types that have a poor survival, although this has not been shown for prostate cancer [21].

Staging investigations are especially valuable for men who are candidates for curative treatment. For this reason it is important to rule out metastatic disease in high-risk men, with systemic imaging, to ensure that radical local treatment is indicated. Therefore, men with low-risk disease are unlikely to undergo these staging investigations given that metastases are unlikely. This is supported in the 2014 NICE guidance where men with low-risk disease should not be routinely offered bone scans [2]. Men with intermediate-risk prostate cancer may undergo a bone scan but again, the likelihood of this being positive is low and negative results are more likely to be missing than positive ones. Our specific clinical assumptions follow these clinical patterns and so we imputed missing N- and M-stage data only in low/intermediate-risk men. Our approach is supported by a study of Swedish cancer registry data which showed that men with incomplete staging tended to be of low- or intermediate-risk [24].

4.5. Strengths and limitations

Strengths of this population-based study include the high volume of patients included (139,807). English cancer registry data includes all NHS hospitals in England which ensures that our findings are nationally generalisable. Second, we used a validated method to identify co-morbidity data in linked hospital administrative data which provided reliable identification of co-morbidity [16].

An important limitation of our staging algorithm is that we did not have access to PSA data in our patient group and we could not sub-classify tumour stage T2 into T2a, T2b or T2c. As a result, we could not use the full D'Amico classification [29]. PSA is now included in the data

Table 3
Results of Cox regression analysis for exploring the mortality of patients according to N-stage (N0/NX/N1) or M-stage (M0/MX/M1) for the patient cohorts applicable to clinical assumptions 1–3.

Clinical assumption	N (%)	Unadjusted hazard ratio (95% CI)	Adjusted hazard ratio[a] (95% CI)
1: Complete N-stage	66,576 (100)		
M0	54,788 (82.3)	1	1
MX	6950 (10.4)	1.08 (0.99–1.19)	1.39 (1.27–1.53)
M1	4838 (7.3)	9.40 (8.88–9.94)	7.07 (6.68–7.50)
2: Low/intermediate-risk	40,664 (100)		
N0	25,705 (63.2)	1	1
NX	14,594 (35.9)	1.50 (1.36–1.65)	1.26 (1.13–1.40)
N1	365 (0.9)	3.53 (2.57–4.85)	3.02 (2.20–4.16)
3: Low/intermediate-risk	40,664 (100)		
M0	25,247 (62.1)	1	1
MX	15,021 (13.9)	1.21 (1.10–1.34)	1.21 (1.09–1.34)
M1	396 (1.0)	7.77 (6.30–9.57)	5.40 (4.37–6.66)

[a] Adjusted for age, RCS Charlson score, socio-economic deprivation, year of diagnosis and English Cancer Registry region.

Table 4
Results of Cox regression analysis for exploring the mortality of patients according to N-stage (N0/NX/N1) or M-stage (M0/MX/M1) following the creation of 10 complete data sets using multiple imputation by chained equations for the patient cohorts defined by clinical assumptions 1–3.[a]

Clinical assumption	%	Unadjusted hazard ratio (95% CI)	Adjusted hazard ratio[b] (95% CI)
1: Complete N-stage	100		
M0	82.3	1	1
MX imputed to M0	9.1	0.85 (0.76–0.96)	1.13 (1.00–1.28)
MX imputed to M1	1.3	2.78 (2.30–3.35)	2.53 (2.07–3.09)
M1	7.3	9.40 (8.88–9.94)	7.11 (6.71–7.53)
2: Low/intermediate-risk	100		
N0	62.1	1	1
NX imputed to N0	35.1	1.48 (1.34–1.64)	1.25 (1.12–1.39)
NX imputed to N1	0.8	2.06 (1.16–3.64)	1.68 (0.96–2.94)
N1	0.9	3.53 (2.57–4.85)	3.03 (2.21–4.17)
3: Low/intermediate-risk	100		
M0	62.1	1	1
MX imputed to M0	35.6	1.18 (1.06–1.31)	1.20 (1.08–1.34)
MX imputed to M1	1.3	2.06 (1.24–3.43)	1.23 (0.75–2.04)
M1	1.0	7.77 (6.30–9.57)	5.41 (4.38–6.69)

[a] See the methods section for a further explanation of the multiple imputation method used and how the resultant ten data sets were combined for the Cox regression results.
[b] Adjusted for age, RCS Charlson score, socio-economic deprivation, year of diagnosis and English Cancer Registry region.

that is being recorded in the English cancer registry which will allow for a more accurate risk stratification [17].

4.6. Clinical implication

If missing staging data is not handled correctly, it can produce biased results. Our study validates the use of specific clinical assumptions to improve staging completeness so that subsequent work using cancer registry data is rendered more reliable. A key advantage of these assumptions over statistical methods is their simplicity and easy application. The clinical assumptions impute just one value for each missing N or M variable, whereas multiple imputation produces a distribution of values (across several data sets). This means that with our clinical rules local hospitals can summarise and analyse their data easily without the need for statistical software or expertise.

One could also use this method to account for missing data using clinically relevant information, followed by statistical imputation as a second step to account for the remainder. Although we have only shown this approach is as appropriate for prostate cancer it may also be applicable to other cancers such as breast, bladder or melanoma. Prostate cancer is unique in that full staging investigations are not always warranted, and treatment not always required, which may limit its use in other cancer types. Specific assumptions would need to be tailored to each cancer type given the differences in their diagnostic pathways.

5. Conclusions

National cancer registries are important data sources for research and healthcare service evaluation. How to handle missing data is therefore of particular importance as historic prostate cancer data are currently being used with relatively high rates of missing cancer stage. Our clinical imputation approach can be an important first step to improve the completeness of cancer stage data, prior to employing statistical imputation techniques, or if statistical methods are not feasible.

Appendix A

See Table A1.

Funding

M.G.P. was partly supported by the NHS National Institute for Health Research through an Academic Clinical Fellowship (ACF-2014-20-002). H.P. was supported by the University College London Hospitals/University College London Comprehensive Biomedical Research Centre. J.v.d.M. was partly supported by the National Institute for Health Research Collaboration for Leadership in Applied Health Research and Care North Thames at Bart's Health NHS Trust. M.G.P., A.S., T.C., S.C., J.N., A.A., N.W.C., H.P., and J.v.d.M. are members of the Project Team of the National Prostate Cancer Audit (www.npca.org.uk) funded by the Healthcare Quality Improvement Partnership (www.hqip.org.uk). The views expressed in this article are those of the authors and not necessarily those of the NHS, the National Institute for Health Research or the Department of Health and Social Care.

Author's contribution

M.G.P. designed the work, analysed and interpreted the data, drafted the article and approved the final version to be published. A.S. designed the work, analysed and interpreted the data, provided critical revision and approved the final version to be published. T.E.C. analysed and interpreted the data, provided critical revision and approved the final version to be published. S.C. analysed and interpreted the data, provided critical revision and approved the final version to be published. J.N. provided critical revision and approved the final version to be published. A.A. designed the work, analysed and interpreted the data, provided critical revision and approved the final version to be published. N.W.C. provided critical revision and approved the final version to be published. H.P. provided critical revision and approved the final version to be published. J.v.d.M. designed the work, analysed and interpreted the data, contributed to the drafting of the article, and approved the final version to be published.

Conflict of interest

J.v.d.M. reports a contract with the Healthcare Quality Improvement Partnership for the provision of the National Prostate Cancer Audit (www.npca.org.uk) funded by the Healthcare Quality Improvement Partnership (www.hqip.org.uk).

H.P. has attended and received honoraria for advisory boards, travel expenses to medical meetings, and served as a consultant for AstraZeneca, Astellas, Janssen, Sanofi Aventis, Takeda, Ipsen, Ferring, Sandoz, and Novartis.

N.W.C. has attended and received honoraria for advisory boards, travel expenses to medical meetings, and served as a consultant for AstraZeneca, Astellas, Bayer, Janssen, Sanofi Aventis, Takeda, Ipsen and Ferring.

Acknowledgments

We thank NHS staff for their support of collecting the clinical data, the National Cancer Registration and Analysis Service (www.ncras.nhs.uk) for providing cancer registry data and NHS Digital (www.digital.nhs.uk) for providing Hospital Episode Statistics; re-used with permission of NHS Digital; all rights reserved.

Table A1
Completeness of individual TNM staging items with and without clinical imputation.

			TNM completeness				
			Without assumptions			With assumptions	
	N	%	T n (%)	N n (%)	M n (%)	N n (%)	M n (%)
Total	139,807	100	95,745 (68.5)	66,576 (47.6)	75,218 (53.8)	81,170 (58.1)	93,909 (67.2)
Year of diagnosis							
2010	33,701	24.1	15,192 (45.1)	6976 (20.7)	8090 (26.4)	10,235 (30.4)	12,903 (38.3)
2011	33,463	23.9	20,099 (60.1)	12,836 (38.4)	13,696 (40.9)	16,401 (49.0)	18,754 (56.0)
2012	35,225	25.2	29,361 (83.4)	22,398 (63.6)	25,481 (72.3)	26,010 (73.8)	29,990 (85.1)
2013	37,418	26.8	31,093 (83.1)	24,366 (65.1)	27,132 (72.5)	28,523 (76.2)	32,262 (86.2)
Age group (yrs)							
≤60	15,321	11.0	11,808 (77.1)	8653 (56.5)	8213 (53.6)	10,606 (69.2)	11,165 (72.9)
60–69	47,923	34.3	36,296 (75.7)	26,815 (56.0)	26,789 (55.9)	32,435 (67.7)	34,799 (72.6)
70–79	51,063	36.5	35,189 (68.9)	24,612 (48.2)	28,154 (55.1)	29,941 (58.6)	34,400 (67.4)
80–89	22,435	16.1	11,514 (51.3)	6037 (26.9)	10,834 (48.2)	7673 (34.2)	12,257 (54.6)
≥90	3,065	2.2	938 (30.6)	459 (15.0)	1228 (40.1)	515 (16.8)	1288 (42.0)
Number of co-morbidities (RCS Charlson score)							
0	109,470	78.3	77,972 (71.2)	54,486 (49.8)	59,107 (53.4)	6644 (60.7)	74,637 (68.2)
1	21,007	15.0	13,087 (62.3)	9006 (42.9)	11,301 (53.8)	10,959 (52.2)	13,661 (65.0)
> 2	9330	6.7	4686 (50.2)	3084 (33.1)	4810 (51.6)	3767 (40.4)	5611 (60.1)
Socio-economic deprivation							
1 (least deprived)	34,259	24.5	23,772 (69.4)	15,947 (46.6)	17,686 (51.6)	20,082 (58.6)	22,983 (67.1)
2	33,669	24.1	23,683 (70.3)	16,271 (48.2)	18,383 (54.6)	20,040 (59.5)	23,100 (68.6)
3	29,533	21.1	20,458 (69.3)	14,269 (48.3)	16,244 (55.0)	17,381 (58.9)	20,183 (68.3)
4	23,462	16.8	15,533 (66.2)	11,018 (47.0)	12,608 (53.7)	13,181 (56.2)	15,363 (65.5)
5 (most deprived)	18,884	13.5	12,299 (65.1)	9071 (48.0)	10,297 (54.5)	10,486 (55.5)	12,280 (65.0)

References

[1] J. Shoag, C.E. Barbie, Clinical variability and molecular heterogeneity in prostate cancer, Asian J. Androl. 18 (4) (2016) 543–548.
[2] National Institute for Health and Care Excellence, Prostate Cancer: Diagnosis and Management, Available at: (2014) (Accessed 16 February 2018), http://www.nice.org.uk/guidance/CG175.
[3] M. Sogaard, M. Olsen, Quality of cancer registry data: completeness of TNM staging and potential implications, Clin. Epidemiol. 4 (Suppl. 2) (2012) 1–3.
[4] U. Nur, L.G. Shack, B. Rachet, J.R. Carpenter, M.P. Coleman, Modelling relative survival in the presence of incomplete data: a tutorial, Int. J. Epidemiol. 39 (1) (2010) 118–128.
[5] G. Gandaglia, F. Bray, M.R. Cooperberg, et al., Prostate cancer registries: current status and future directions, Eur. Urol. 25 (5) (2015) 979–986.
[6] S.H. Giordano, Y.F. Kuo, Z. Duan, G.N. Hortobagyi, J. Freeman, J.S. Goodwin, Limits of observational data in determining outcomes from cancer therapy, Cancer 112 (11) (2008) 2456–2466.
[7] D.F. Penson, P.C. Albertsen, Lessons learnt about early prostate cancer from large scale databases: population-based pearls of wisdom, Surg. Oncol. 11 (1–2) (2002) 3–11.
[8] Q. Luo, X.Q. Yu, C. Cooke-Yarborough, D.P. Smith, D.L. O'Connell, Characteristics of cases with unknown stage prostate cancer in a population-based cancer registry, Cancer Epidemiol. 37 (6) (2013) 813–819.
[9] M. Nguyen-Nielsen, T. Froslev, S. Friis, M. Borre, N. Harving, M. Sogaard, Completeness of prostate cancer staging in the Danish cancer registry, 2004–2009, Clin. Epidemiol. 4 (Suppl. 2) (2012) 17–23.
[10] N. Eisemann, A. Waldmann, A. Katalinic, Imputation of missing values of tumour stage in population-based cancer registration, BMC Med. Res. Methodol. 11 (2011) 129.
[11] A. Marshall, D.G. Altman, P. Royston, R.L. Holder, Comparison of techniques for handling missing covariate data within prognostic modelling studies: a simulation study, BMC Med. Res. Methodol. 10 (2010) 7.
[12] M.T. Redaniel, A. Laudico, M.R. Mirasol-Lumague, A. Gondos, G. Uy, H. Brenner, Inter-country and ethnic variation in colorectal cancer survival: comparisons between a Philippine population, Filipino-Americans and Caucasians, BMC Cancer 10 (2010) 100.
[13] Q. Luo, S. Egger, X.Q. Yu, D.P. Smith, D.L. O'Connell, Validity of using multiple imputation for "unknown" stage at diagnosis in population-based cancer registry data, PLoS One 12 (6) (2017) e0180033.
[14] J.A. Sterne, I.R. White, J.B. Carlin, et al., Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls, BMJ 338 (2009) b2393.
[15] World Health Organisation, ICD-10 International Statistical Classification of Disease and Related Health Problems, Available at: (2016) (Accessed 16 February 2018), http://apps.who.int/classifications/icd10/browse/2016/en.
[16] J.N. Armitage, J.H. van der Meulen, Identifying co-morbidity in surgical patients using administrative data with the Royal College of Surgeons Charlson Score, Br. J. Surg. 97 (5) (2010) 772–781.
[17] National Prostate Cancer Audit, Third Year Annual Report - Results of the NPCA Prospective Audit and Patient Survey, Available at: (2016) (Accessed 16 February 2018), http://www.npca.org.uk/annual-report-2016/.
[18] I.R. White, P. Royston, A.M. Wood, Multiple imputation using chained equations: issues and guidance for practice, Stat. Med. 30 (4) (2011) 377–399.
[19] J. Gurney, D. Sarfati, J. Stanley, et al., Unstaged cancer in a population-based registry: prevalence, predictors and patient prognosis, Cancer Epidemiol. 37 (4) (2013) 498–504.
[20] A.C. Klassen, F. Curriero, M. Kulldorff, A.J. Alberg, E.A. Platz, S.T. Neloms, Missing stage and grade in Maryland prostate cancer surveillance data, 1992–1997, Am. J. Prev. Med. 30 (2 Suppl) (2006) S77–87.
[21] R.M. Merrill, A. Sloan, A.E. Anderson, K. Ryker, Unstaged cancer in the United States: a population-based study, BMC Cancer 11 (2011) 402.
[22] S.P. Elliott, D.P. Johnson, S.L. Jarosek, B.R. Konety, O.O. Adejoro, B.A. Virnig, Bias due to missing SEER data in D'Amico risk stratification of prostate cancer, J. Urol. 187 (6) (2012) 2026–2031.
[23] M. Radespiel-Troger, K. Geiss, A. Gartig-Daugs, M. Meyer, Registration completeness and tumour-node-metastases category distribution in prostatic and testicular cancer in Bavaria, Eur. J. Cancer Prev. 17 (3) (2008) 230–237.
[24] K. Tomic, M. Westerberg, D. Robinson, H. Garmo, P. Stattin, Proportion and characteristics of men with unknown risk category in the National Prostate Cancer register of Sweden, Acta Oncol. 55 (12) (2016) 1461–1466.
[25] L. Barlow, K. Westergren, L. Holmberg, M. Talback, The completeness of the Swedish Cancer register: a sample survey for year 1998, Acta Oncol. 48 (1) (2009) 27–33.
[26] C.J. Bradley, J.P. Clement, C. Lin, Absence of cancer diagnosis and treatment in elderly Medicaid-insured nursing home residents, J. Natl. Cancer Inst. 100 (1) (2008) 21–31.
[27] S.M. Koroukian, F. Xu, H. Beaird, M. Diaz, P. Murray, J.H. Rose, Complexity of care needs and unstaged cancer in elders: a population-based study, Cancer Detect. Prev. 31 (3) (2007) 199–206.
[28] J.L. Worthington, S.M. Koroukian, G.S. Cooper, Examining the characteristics of unstaged colon and rectal cancer cases, Cancer Detect. Prev. 32 (3) (2008) 251–258.
[29] A.V. D'Amico, R. Whittington, S.B. Malkowicz, et al., Biochemical outcome after radical prostatectomy, external beam radiation therapy, or interstitial radiation therapy for clinically localized prostate cancer, JAMA 280 (11) (1998) 969–974.