

Workplace Project Portfolio (WPP) Masters in Biostatistics

BSTA5021

BSTA 5022

Project 1: Bladder and non-bladder urinary cancers: examining patterns and risk factors for second cancers using data from the New South Wales Central Cancer Registry (Australia)

Project 2: Multiple Imputation to address a data artefact for the degree-of-spread variable in the NSW CCR for the period 1993 – 1998: Lung Cancer as a test case

Heidi Welberry

November 2011

Contents

PART A: Preface.....	2
PART B: Project 1 – Bladder and non-bladder urinary cancers: examining patterns and risk factors for second cancers using data from the New South Wales Central Cancer Registry (Australia).....	11
PART C: Project 2 – Multiple Imputation to address a data artefact for the degree-of-spread variable in the NSW CCR for the period 1993 – 1998: Lung Cancer as a test case.....	37

PART A: Preface

My Role

I completed both projects through a part-time secondment to the Monitoring, Evaluation and Research Unit (MERU) at the Cancer Institute NSW, between June 2010 and April 2011. This arrangement was formally endorsed by the Chief Cancer Officer. Both projects were supervised by Dr Stephen Morrell (statistical supervisor), and Ms Deborah Baker, Manager of MERU.

Both projects were taken from the MERU workplan, using data from the NSW Central Cancer Registry (NSW CCR) and were not familiar to me beforehand. While the projects both used the same dataset, they were not related.

The first project built upon previous work within MERU to explore patterns of second cancer occurrence for urinary cancers. The intention was to increase our knowledge regarding bladder cancers occurring as secondary primary cancers and publish the results as a journal article. A draft manuscript was completed and is currently in the process of review within the CINSW prior to being finalised and submitted.

The second project was focussed on a particular issue that had been identified with coding of the degree of spread variable within the NSW CCR. The aim was to explore, test and report on a technique to address the issue. A report summarising the findings was completed and submitted to MERU. This focussed on lung cancer as a test case and the wider applicability of the technique to other cancers will now be explored further.

This was my first exposure to the Central Cancer Registry Dataset. I completed both projects independently, seeking input where required to understand the dataset, coding, existing SAS programs and formats and discuss issues as they arose.

Reflections on Learning

Communication

Both projects helped me develop communication skills in two key areas. Firstly, throughout the course of the projects, it was essential to negotiate time and communicate with key individuals to gain a thorough understanding of the database and existing resources such as supporting documentation, SAS codes and formats. For both projects, I drew heavily upon the

expertise of other biostatistical staff as well as key staff involved in the coding and management of the NSW CCR. Clear communication of my queries, including the background and context to the project was essential in gaining the information required.

Secondly, both projects required me to communicate complex statistical concepts and results in written and graphical form, as well as verbally when presenting findings to my supervisors and others within the institute. One of the biggest challenges was doing so in a succinct manner and isolating the key results that most directly addressed the questions of interest.

Work Patterns/Planning

Given my part time working arrangements on these projects, setting work patterns and planning ahead were essential in allowing me to balance my time on the projects with my other quite different work requirements.

Completing both projects within reasonably tight timeframes required self-discipline, motivation as well as a flexible approach. I set deadlines for sections of work as well as overall completion of the projects. However, I found that both projects included unanticipated complexities due to coding which meant that I had to revisit the scope of the projects in consultation with my supervisors and revise my approach and timeframes where necessary.

Both projects have helped me to realise that working with administrative data often requires compromise and more questions can often be raised than answers generated. Within project 1, it was only after gaining a more thorough knowledge of the dataset and the limitations of some variables that I felt that I could adequately design specific analyses that would address the research questions. This meant that the project evolved somewhat from start to finish. Similarly within project 2, understanding the complex relationships between variables within the NSW CCR was essential before being able to make informed decisions regarding the most appropriate approach to be taken. Both projects have generated many new questions and it was difficult at times to contain the scope of the projects within a manageable limit.

Statistical Principles, methods and computing

The Central Cancer Registry was the main dataset used within both projects. This dataset is population-based and aims to capture all registrations of invasive cancers for NSW. The importance of this dataset is in monitoring trends in cancer registrations over time and as such it requires a census of cancer registrations rather than a sample and requires rigid and consistently applied data collection procedures.

Both projects gave me a very good insight into the importance of understanding the NSW CCR data collection and coding practices in detail. While some variables on the NSW CCR are reasonably straight-forward, others such as coding site of cancer, histology of cancer and degree of spread are governed by complex international classifications and coding rules. Additionally, the information provided from notifying sites for these variables is not always of sufficient quality to be able to code as accurately as would be desired. Communicating with experts from both a pathology and coding perspective was essential to understanding and using these variables in a valid and informative manner.

It is also essential that inferences related to trends in cancer registrations are made with reference to trends in the underlying population. As such, population estimates from the Australian Bureau of Statistics (ABS) are an important component of any population-based cancer estimates.

Project 1

Project 1 examined patterns of increased risk of a second cancer among those with a prior diagnosis of kidney cancer compared to the general population. The ability to draw a comparison between this sub-group and the underlying population was only possible because of the population based approach of the NSW CCR.

Another important epidemiological principle for this project was the concept of “person years of observation”. When monitoring a population for the occurrence of a particular disease/event it is very important to consider the time period for which they were at risk of the event. For this project the event of interest was a second cancer and people were at risk from the time they entered the registry (at diagnosis of a first cancer) to the time they died, were diagnosed with a second cancer or until the end of the follow-up period. It was essential that time of death could be taken into account as this was highly likely to be related to the type of cancer with which they were initially diagnosed.

Two main statistical methods were employed in this project: (i) the calculation of Standardised Incidence Ratios (SIRs); and (ii) Survival Analysis. Both were chosen because of their ability to account for biases in person years of observation due to data censoring. However, the first technique allowed examination of patterns of second cancers and to compare these with the underlying population, while the second was used to examine predictors of second cancer occurrence within a sub-group of people who had a specific cancer diagnosis.

I had come across the concept of SIRs previously within epidemiology courses undertaken and within journal articles and understood the principle of the calculation, although I had not previously undertaken such calculations using the retrospective cohort approach. I had access to a SAS program that had been previously created within MERU to calculate SIRs on CCR data, so did not need to start from scratch. However, in order to make any modifications to this program I needed to understand in a reasonable amount of detail how it worked.

In undertaking the survival analysis component, I drew heavily upon the survival analysis unit I had completed in first semester 2010. There were some major differences though when undertaking this type of analysis on a large-scale administrative dataset. The potential length of survival time was much longer than any previously encountered, and because the CCR has now been operating for almost three decades, the potential issues with data inconsistency on some variables were much more complex than I anticipated. A major challenge with this part of the analysis was maintaining a connection to a relevant clinical or epidemiological question while balancing this with the limitations of the data.

Comparing these two techniques provided a very good insight into the benefits and limitations of each. While SIRs provided a very good overview of elevated risk of a second cancer compared to the population level this technique could not be used to examine the significance of covariates in predicting second cancers within the population sub-group. A clear example of this was the relationship between sex and occurrence of bladder cancers following initial upper urinary tract cancers (eg. of the renal pelvis). SIRs showed the occurrence of bladder cancers to be much more markedly elevated for females compared to males. However, survival analysis showed that there was no difference between males and females in the rate of second cancer occurrence. SIRs were influenced by the underlying sex-specific population incidence of both first and second cancers whereas survival analysis only took account of sex-specific incidence of second cancers within the population of interest. This was an important lesson in drawing conclusions from each technique.

One of the major challenges I faced when commencing this project was using SAS for all analyses. While I had used SAS previously, including for multivariate analyses, there had been a period of at least two years where my use of SAS had been limited with use of other software packages instead during this time such as STATA. This was particularly an issue in conducting survival analysis which I had only ever previously undertaken in STATA and in which the BCA Survival Analysis unit I had completed was entirely based. I could have taken the option of installing STATA and using this, but chose to persist with SAS given it is so widely used at CINSW and there were existing SAS resources for CCR data such as programs and formats that I

could draw upon. While SAS provided powerful options for the programming required for SIR calculations and also survival analysis, I found generating useful data and plots to assess model diagnostics to be less elegant and more time consuming within SAS compared to STATA.

Project 2

Of particular importance in project 2 were the principles of random and non-random variation and, in terms of missing data, the differences between Missing Completely at Random (MCAR), Missing at Random (MAR) and Missing Not at Random (MNAR). The issue under investigation involved data with a partially known missing data mechanism. The missing data was within a discrete period of time; however it was also related to the notification method used.

It was evident that for many variables being investigated, including the main variable of interest *degree of spread*, there had been non-random variation in coding and notification practices over time. The use of imputation to address missing data assumes the data are not MNAR, therefore considerable deliberation was required to choose an approach that best reflected this assumption based on the known relationships between variables. However, it was also important to clearly articulate where assumptions had been made.

In building a valid logistic model, it was important to consider the predictive power of the model. To do so, I drew upon principles of test evaluation which consider the sensitivity and specificity of a test in predicting an outcome. This provided a useful basis for assessing in what contexts the imputation process may be valid and where caution may need to be exercised in using this approach.

The main statistical methods employed within this project were focussed on two areas: (i) categorical data analysis, using chi-square tests and logistic regression; and (ii) data imputation using Proc MI in SAS. Additionally, to validate the effect of the imputation process, modelling of imputed results against 'test' cases was conducted, and Kaplan-Meier survival curves were constructed.

The first part of this project required building a logistic regression model to predict whether a lung cancer case had localised or unknown degree of spread. For the model building process, I drew upon courses such as Categorical Data and Generalized Linear Models. However, in most courses model building focussed on investigating relationships between covariates and an outcome variable rather than on predictive power. The process of building a model, given the specific aim and data limitations within this project, was a balance between gaining predictive

power and gaining consistency across the two variables that were related to the missing data mechanism: time period and notification method. This process reinforced the opinion that model building is not an exact science and that many different choices could be made. The end result was achieved by finding a model that could be defended as appropriate given the assumptions, but also provided adequate prediction. This was a much more difficult process than I originally anticipated.

While missing data were acknowledged in many of the BCA courses as being problematic in a real world setting, no course specifically addressed the issue. In practice, there is often no ideal solution to addressing the issue and instead the best option must be chosen based on the extent of the problem and the type of mechanism causing the missing data. This project highlighted the importance of getting the data collection process right in the first place so as to avoid complex missing data patterns. Multiple imputation appeared a useful tool to address a complex missing data problem, but it was clear in the current project that the usefulness of this approach is dependent on understanding the missing data pattern and being able to construct a valid predictive model.

SAS was used for all analyses with graphs produced either within SAS or via Excel. A variety of different procedures were used including Proc Logistic, Proc MI, Proc MIanalyze and Proc Lifetest. Much of the investigation of potential predictors was quite repetitive and one of the main time-saving mechanisms I developed was to output files directly to excel using the output delivery system (ods) within SAS. I also explored this approach for directly outputting graphs required for the report, but in the end found it quicker and more effective to produce the graphs within excel from the outputted data using a graph template. I would like to develop further my knowledge and expertise with using the output delivery system, as can see many advantages, both in terms of saving time and setting up efficient and standardised processes for running routine or repetitive data analysis.

The Proc MI and Proc MIanalyze procedures were new to me for this project and I relied heavily on the SAS documentation to understand how to correctly apply these procedures. The missing data mechanism within this project meant that I had to think creatively about how to adapt the procedure to use prototype cases from one period of time to predict values for missing data in another period.

Teamwork

Communication with other team members

To complete both projects I drew upon the expertise of multiple other people within MERU and also within the Central Cancer Registry. I built key relationships with the biostatistics team to ensure correct understanding and use of databases, SAS programs and SAS formats. In particular, for project 1 this included the understanding of the SAS program for creation of SIRs. I found the most efficient way to work with this team was to outline my queries in writing via email and then book meeting times when necessary to go through processes or code in more detail.

I also liaised with the coding manager and consultant pathologist advising on coding for the Central Cancer Registry. For project 1, communication with the consultant pathologist in particular was essential in gaining an understanding of the cell types and histological sub-types of cancers and also for assessing the relevance of particular questions to the clinic. For project 2, the coding manager was an essential resource to understanding current and past coding practices for the degree-of-spread variable. I was extremely fortunate that she has been involved with the NSW CCR for many years, including during the 1990's when the data collection issue for degree of spread occurred.

Communication on progress with my supervisors was conducted mostly by email with regular meetings scheduled every two-three weeks. Again the most effective means of communicating was to ensure material was sent in advance and flag any particular questions for discussion.

Negotiating Roles and Responsibilities

As I was the only resource for both projects, there was little need to negotiate roles and responsibilities for the key tasks within the project. However, as the projects proceeded there was continued communication with my supervisors on the progress made and where input was required from others. This input was mostly in the form of expertise/advice or feedback. For project 2, I required assistance with correct extraction and creation of a variable that would provide me with the method of notification. This required negotiating some time with one of the biostatisticians during a very busy period.

As I was only seconded part time to work on this project, I also needed to negotiate carefully with my Director as well as supervisors to ensure everyone was comfortable with the amount of time I would be allocating to the projects and other tasks. This continued throughout the projects as needed.

Working within timelines

I initially set my own timelines for the projects in consultation with my supervisors. I worked on the two projects in sequence, setting a deadline for project 1 in November 2010 and a deadline for project 2 in April 2011. The tight timeframe for project 2 was necessitated due to my personal circumstances, as I discovered I was expecting a baby in May 2011. Further consultation and discussion regarding project 1 has continued since November, with some further revisions planned for the manuscript prior to submitting this to a journal for publishing.

The main mechanism for adhering to timelines was preparing a plan in advance of what milestones would be reached by each supervisory meeting. Timeframes required re-setting slightly and the scope of the projects required adjusting once work commenced due to unanticipated complexities in the coding of data. For project 1, this meant that more time was spent up-front understanding the issues so as to avoid incorrect or misleading use of the data. So as to still complete within the timeframe and provide a better focus for the project, it was decided that some analyses initially included in the project scope were better left out of the final manuscript. These will be followed up separately. For project 2, after initial exploration of the data and relationships between variables it was decided that some extra input from one of the other biostatisticians was required to create a variable that indicated the notification method for cases as this was a key aspect of the missing data mechanism. As it was a very busy period for the biostatistics team, this meant timelines had to be adjusted to fit with their workload.

One of the major challenges in keeping the projects on time was that as well as my own part-time input into the project, one of my supervisors was only available two days per week and the consultant pathologist was only available one day a week. This meant that planning of meetings had to be made well in advance to ensure input from all people.

Helping others to understand statistical issues

As I worked independently on both projects my main focus was on ensuring that I understood the statistical issues and could communicate these to my supervisors in a clear and informative manner. Prior to meetings, I wrote a summary of the issues and my proposed response to the issues so that we could then discuss the alternatives in an informed manner. Additionally, when meeting with others, such as the coding manager or other biostatisticians, I tried to summarise the problem I wanted to discuss in a very concise manner. I found in many

circumstances being able to graphically represent the problem was the most effective way to do this, particularly for discussions with the non-biostatisticians.

In preparing the manuscript and report that resulted from these projects, I focussed on trying to clearly communicate the statistical issues involved and provide a clear explanation for the approach taken. For example, in the report for Project 2, I included a discussion regarding the background to the NSW CCR and the different types of missing data to try to provide some context to the problem being addressed.

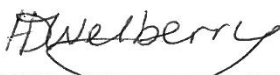
Ethical Considerations


At the beginning of these projects, I ensured I was familiar with the NHMRC National Statement on Ethical Conduct on Research Involving Humans and also the NSW Health Act and Privacy Act to understand the context in which NSW Cancer Registry data is collected and how this fits within the National Statement. Both projects utilised NSW CCR data only and did not require any identifiable variables or data linkage. They were carried out within the remit of the core functions of the Monitoring Evaluation and Research Unit at the CINSW, and as such, were determined to be low risk projects that did not require ethical review.

However, I was still conscious of my obligations in maintaining the confidentiality and privacy of the data. This included ensuring that data security was considered by carrying out all analyses within password protected secure servers at the CINSW. It also included consideration of confidentiality in reporting. As some of the combinations of cancers under investigation had few observations, the decision was made to report any cells with values less than 5 as approximate only (ie. "<5").

Throughout the project I was also aware of my professional responsibility to carry out accurate analyses and provide correct and relevant interpretation. This was especially important after uncovering certain limitations with the coding of the data and I made it a priority to understand these limitations in detail, adjust the analyses accordingly, and report the limitations of the study.

PART B: Project 1 – Bladder and non-bladder urinary cancers: examining patterns and risk factors for second cancers using data from the New South Wales Central Cancer Registry (Australia)

Location and Dates:	Cancer Institute NSW, Sydney, June-November 2010
Context:	This project was completed as part of the approved program of work within the Monitoring Evaluation and Research Unit at the Cancer Institute NSW. Heidi was seconded to MERU from another area of the CINSW to work on this project on a part-time basis. This project extended previous work completed within MERU assessing the excess risk of second primary cancers following an initial primary cancer diagnosis.
Contribution of student:	Heidi completed all review of literature, design of study, analyses, and write-up/presentation of results. The calculation of SIRs utilized an existing SAS program created within the CINSW. The Survival Analysis component was completed independently.
Statistical issues involved:	<ul style="list-style-type: none"> • Calculation of Standardised Incidence Ratios (SIRs) • Kaplan-Meier survival analysis • Cox Proportional Hazards Regression
Declaration:	I declare that I have undertaken this project independently and have not submitted this work for previous academic credit.
Signed:	

Supervisors Name:	Dr Stephen Morrell (Co-supervisor with Ms Deborah Baker)
Statement:	This is to state that Heidi Welberry has conducted the statistical analyses, as outlined above, and writing for this project independently and in a very competent manner. After reading the relevant literature, Heidi has also conceived the research questions for the project in terms of the relevant issues. Consequently, the work has a high probability of being publishable.
Signed:	

Bladder and non-bladder urinary cancers: examining patterns and risk factors for second cancers using data from the New South Wales Central Cancer Registry (Australia)

Introduction

Study of the incidence of second primary cancers can be informative in identifying cancers with common etiologic factors or cancers that may arise as a consequence of treatment for the initial primary cancer. Following kidney cancer, the observation of excess invasive cancers of the bladder previously has been reported using data from the NSW Central Cancer Registry for the period 1972-1991 (1). McCredie *et al.* reported an elevated risk of invasive bladder cancer following cancer of the renal parenchyma in women only (Risk Ratio (RR)=3.4, 95% CI=1.1-8.0), and an elevated risk of invasive bladder cancer following cancer of the renal pelvis in both men (RR=8.7, 95% CI=5.4-13) and women (RR=39, 95% CI=26-56). They postulated that the pattern of excess cancers following cancer of the renal pelvis supported tobacco as a common risk factor, but not for cancer of the renal parenchyma. They suggested that the increased elevation of risk of bladder cancer in women reflected the high incidence of analgesic-associated disease from use of products containing Phenacetin.

The renal parenchyma and the renal pelvis have distinctly different morphological features. The renal parenchyma comprises nephrons, the functional tissue of the kidney, and cancers occurring in this region are most commonly renal cell carcinoma. The renal pelvis acts as the funnel for urine flowing to the ureter and comprises urothelial tissue. Urothelial tissue (the urothelium) covers the surface of the urinary tract from the renal pelvis through the ureter and the bladder to the proximal urethra. The urothelium is characterized by transitional cells, and in Western countries more than 90% of cancers of these organs are transitional cell cancers (2). Another characteristic of urothelial tumors is that they are frequently multifocal in nature, commonly occurring either synchronously or asynchronously in different regions of the urothelium.

There have been two main hypotheses put forward to explain the multifocal nature of urothelial cancers. The concept of field cancerisation first proposed in 1953 by Slaughter *et al.* suggests that the entire urothelium is exposed to a common risk factor putting the entire 'field' of the urothelium at risk of developing tumours (3). These tumours subsequently develop independently. Alternatively, a more recent theory of intraluminal 'seeding' has been proposed. This suggests that cells from a single tumour or lesion can dislodge and implant at another site. Molecular studies have supported the 'seeding' hypothesis by showing identical mutations in tumours from multiple locations (4). A recent review suggests that the seeding hypothesis is now well supported but that both mechanisms are likely to occur (2).

Two distinct questions can be posed: does the pattern of second urinary cancers support the theory of field cancerisation and/or intraluminal seeding? And what factors are predictive of the rate of second cancer occurrence within the urinary tract? For the first question in particular, the likelihood of a second cancer 'downstream' from the kidney can be compared with the converse, for instance a bladder cancer followed by kidney cancer, to shed light on which mechanism is more likely. NSW Central Cancer registry data provide an opportunity to explore the patterns of urinary cancers occurring together at a population level.

Materials and Methods

All data used were extracted from the NSW Central Cancer Registry. Notifications for invasive cancers are mandatory for pathology laboratories, hospitals and other treatment centres under the NSW Public Health Act 1991. All first and second invasive cases of cancer for an individual were included with third and subsequent cancers excluded from analyses. Two analysis techniques were used to address the research questions: the calculation of standardized incidence ratios to compare observed versus expected numbers of second cancers; and Cox Proportional Hazards regression modelling to investigate predictive factors for second cancer occurrence. SAS version 9.2 was used for all statistical analyses.

Standardised Incidence Ratios

Two sets of analyses were undertaken: i) non-bladder urinary cancers as the first cancer sites (renal parenchyma, renal pelvis, ureter, and urethra) and bladder cancer as the second cancer site; ii) bladder cancer as the first cancer site and non-bladder urinary cancers as the second cancer sites. The first 3 digits of the ICD10 coding system were used to separate organ sites with urethra defined at the 4 digit level. Second cancers diagnosed within three months of the first cancer were excluded.

A cohort model was used where person-years following diagnosis of the first cancer were categorised by single year of age (with an open-ended category from 85 years), sex and calendar year of diagnosis. For each analysis, the event was diagnosis of a second primary cancer of interest (the target cancer). The follow-up period started at 3 months following first diagnosis and was censored at time of death, date of diagnosis of a second primary non-target cancer, or 31 December 2007, whichever occurred first. Expected numbers of cases were calculated based on the age- and sex-specific incidence rate for each calendar year within the follow-up period. Standardised Incidence Ratios (SIRs) were calculated by dividing the observed number of second cancers by the expected number of second cancers and 95%

confidence intervals were calculated based on the Poisson distribution. The expected number of second cancers was based on the incidence of that cancer in the population overall.

Cox Proportional Hazards Regression

All persons diagnosed from 1986 to 2007 with an invasive upper urinary tract cancer (Renal Pelvis: ICD10 code C65 and Ureter ICD10 C66) as a first cancer were included in the cohort. Cases without histological verification at the NSWCCR were excluded. An event was the occurrence of a histologically verified invasive bladder cancer (ICD10 C67) as a second cancer at least 3 months following diagnosis of the first cancer. Individuals were censored at death, diagnosis of a non-bladder cancer or 31st Dec 2007 whichever occurred first. Due to a low number of events in non-transitional cell cancers, the cohort was further restricted to include only Transitional Cell Carcinoma (TCC, histology codes 81203, 81313, 81303, 81223). The cohort comprised 1,700 cases. It should be noted that the inclusion criteria for the Cox regression cohort was much stricter than that used in the SIR analysis, focussing on a shorter time period (starting at 1986 rather than 1972) and a more specifically defined group (only TCC that had been histologically verified). Follow-up began at time of diagnosis of first cancer

Covariates included sex of individual and characteristics of the first cancer diagnosis: age at diagnosis (in years); period of diagnosis (in 4-yearly groups); degree of spread of first cancer at diagnosis (localised; regionalised; distant and unknown); site of first cancer (renal pelvis vs ureter); histology of first cancer (papillary TCC (81303, 81313) vs TCC (81203, 81223)); socio-economic status at diagnosis (approximated using the index of relative advantage and disadvantage based on postcode of residence at time of diagnosis). Proportionality of Hazards was assessed by including time dependent forms of each covariate in the model. A full model was initially tested with all covariates included as predictors of time to event. A final reduced model was constructed by removing non-significant predictors in a step-wise fashion until all variables remaining were significant at $P < 0.05$ level.

Results

Standardised Incidence Ratios

Table 1 presents observed and expected numbers of second cancers by sites of first and second cancer and sex. With the exception of bladder cancers following cancers of the renal parenchyma for men, all standardized incidence ratios (SIRs) were significant. No bladder cancers were observed following cancers of the urethra for women. The largest SIRs observed

were for bladder cancer following cancers of the renal pelvis (63.8; 95% CI: 52.3-77.1) and ureter (80.1; 95% CI: 56.1-111) for women, and urethral cancer following bladder cancer for men (58.3; 95% CI: 38.7-84.2).

For bladder cancers following non-bladder urinary cancers, SIRs were larger for papillary TCC compared to TCC. SIRs trended downwards over time with the SIRs for bladder cancer following Papillary TCC remaining significant for both males and females at more than 10 years of follow-up whereas for bladder following TCC, SIRs at 10+ years dropped to non-significance (Table 2).

All SIRs for combinations of cancers in which the second cancer could be considered 'downstream' of the first were larger than for the corresponding 'upstream' combination of cancers excepting for cancer of the renal parenchyma. Non-overlapping confidence intervals for upstream compared to downstream combinations demonstrate the significance of this relationship with the exception of bladder/urethra for women related to the very low numbers of urethral cancers in females.

Cox Proportional Hazards Regression

1,700 patients were included in the cohort with a median follow-up after diagnosis of an upper urinary tract cancer as a first cancer of 1.8 years. There was a slight over-representation of females, more renal pelvic cancer cases than ureter and most cases had either localised or regionalised spread at time of diagnosis. Of the 1,700 patients, 137 (8.0%) developed an invasive bladder cancer as a second cancer with a median time to second cancer development within this group of 1.0 year.

Of the seven variables investigated as predictors of time to bladder cancer, age group, tumour site and histological sub-type were significant at the univariate level (Table 3). Those with cancer of the renal pelvis developed bladder cancer at a lower rate than those with cancer of the ureter (HR=0.65, p=0.02), and those diagnosed with a papillary TCC (histology code 81303, 81313) were more likely to develop a cancer of the bladder compared to those with TCC (histology code 81203, 81223) (HR=1.58, p=0.01). Those in the two oldest age groups were more likely to develop bladder cancers compared to those in the youngest age group (HR=1.88, p<0.01).

At a multivariate level, after controlling for all other factors, age group at first cancer diagnosis, tumour site and histological sub-type remained as significant as predictors. Results from the full model are presented in Table 4 and results from the reduced final model are presented in Table 5. Based on the final model, those aged 65-74 years at diagnosis experienced second

cancers at almost double the rate of those younger than 65 years (HR=1.91). Those aged 75+ years had a 68% increased rate of second cancers compared to the < 65 age group. The difference between the 65-74 year group and 75+ group was not significant (p=0.36). Initial cancers of the renal pelvis were followed by second bladder cancers at only two thirds the rate of initial ureter cancer. Initial cancers classified as Papillary Transitional Cell Carcinomas (histology code 81303) had a 59% higher rate of second bladder cancers compared to initial cancers classified as Transitional Cell Carcinoma (histology code 81203).

The overall test for non-proportionality of hazards was non-significant (Wald $\chi^2=3.25$, p=0.52).

Table 1: Observed and expected numbers of second cancers: kidney, renal pelvis, ureter and urethral cancers following bladder, and bladder cancers following kidney, renal pelvis, ureter and urethra; by sex, 1972-2007

Sex	First Cancer	Person years of observation	Second cancer	Observed number of second cancers	Expected number of second cancers*	SIR	95% CI	Second cancer up or down-stream	
Male	Renal Parenchyma	102,991		21	22.5	0.93	(0.58-1.43)	Down	
	Renal Pelvis	11,876	Bladder	79	3.8	21.0	(16.6-26.1)	Down	
	Ureter	5,612		35	1.9	18.7	(13.1-26.1)	Down	
	Urethra	863		3	0.2	13.2	(2.70-38.4)	Up	
	Bladder		245,315	Renal Parenchyma	56	41.1	1.36	(1.03-1.77)	Up
				Renal Pelvis	71	5.6	12.6	(9.80-15.9)	Up
				Ureter	46	2.5	18.4	(13.4-24.5)	Up
				Urethra	28	0.5	58.3	(38.7-84.2)	Down
	Female	Renal Parenchyma	66,348		11	5.1	2.14	(1.07-3.83)	Down
		Renal Pelvis	16,222	Bladder	107	1.7	63.8	(52.3-77.1)	Down
Ureter		3,682		36	0.4	80.1	(56.1-111)	Down	
Urethra		287		-	0.0	-	-	Up	
Bladder			88,796	Renal Parenchyma	24	8.9	2.71	(1.74-4.03)	Up
				Renal Pelvis	34	3.0	11.4	(7.90-16.0)	Up
				Ureter	26	0.7	37.1	(24.2-54.4)	Up
				Urethra	3	0.1	39.4	(8.10-115)	Down

Table 2: SIRs by time period following diagnosis and overall for non-bladder urinary as first and bladder as second cancer; by histology of first cancer and sex

First Cancer	(3mo-<5yrs)		(5-<10yrs)		(10+ yrs)		(overall)	
	SIR	95%CI	SIR	95%CI	SIR	95%CI	SIR	95%CI
Female								
Papillary TCC (8130/3)	136	(105, 174)	26.6	(10.7, 54.8)	22.2	(8.9, 45.8)	74.3	(58.7, 92.7)
TCC (8120/3)	113	(85.1, 146)	23.7	(7.7, 55.4)	2.90	(0.10,16.1)	58.8	(45.1, 75.4)
ALL TCC	124	(103, 148)	25.3	(13.1, 44.2)	12.1	(5.2, 23.8)	66.6	(56.0, 78.5)
Male								
Papillary TCC (8130/3)	52.5	(40.1, 60.0)	10.6	(3.88, 23.0)	7.70	(2.8, 16.8)	22.6	(22.6, 36.4)
TCC (8120/3)	26.0	(18.2, 35.9)	7.80	(2.87, 17.0)	1.00	(0.0, 5.40)	9.74	(9.70, 18.1)
ALL TCC	38.0	(30.7, 46.3)	9.00	(4.65, 15.7)	3.80	(1.60,7.90)	16.7	(16.7, 24.3)

Table 3: Demographic and clinical characteristics of first upper urinary tract cancer diagnosis and univariate relationship of covariates to second bladder cancer diagnosis

Variable	Number of observations		Unadjusted Hazard Ratio (95% CI)	P value (Log-rank Test)
	N	%		
Age				
<65yrs	481	28.3	-	
65-74yrs	645	37.9	1.95 (1.26, 3.02)	0.03*
75+yrs	574	33.8	1.79 (1.11, 2.89)	
Sex				
Male	691	40.6	-	1.00
Female	1009	59.4	1.00 (0.71, 1.41)	
Tumour Site				
Renal Pelvis	1305	76.8	-	0.02
Ureter	395	23.2	0.65 (0.46, 0.94)	
Degree of spread				
Localised	737	43.4	-	0.37
Regionalised	644	37.9	0.67 (0.21, 2.14)	
Distant	162	9.5	0.88 (0.59, 1.30)	
Unknown	157	9.2	1.40 (0.85, 2.32)	
Histological Sub-type				
TCC	937	55.1	-	0.01
Papillary TCC	763	44.9	1.58 (1.11, 2.24)	
Period of Diagnosis				
1986-1989	305	17.9	-	0.19*
1990-1994	373	21.9	1.39 (0.79,2.44)	
1995-1998	337	19.8	1.89 (1.08, 3.29)	
1999-2003	393	23.1	1.50 (0.85, 2.66)	
2004-2007	292	17.2	1.14 (0.56, 2.32)	
IRSAD quintile				
Lowest	213	12.5	-	0.39*
Second	269	15.8	1.79 (0.92, 3.47)	
Third	477	28.1	1.40 (0.75, 2.62)	
Fourth	374	22.0	1.35 (0.70, 2.57)	
Highest	360	21.2	1.13 (0.57, 2.23)	
Missing	7	0.4	-	

*indicates a log-rank test for trend, all other values based on log-rank test for difference

Table 4: Cox Proportional Hazards Regression – Full Model: Time from upper urinary tract TCC diagnosis to diagnosis of bladder cancer as a second cancer– sex, age, site of first primary, period of diagnosis, degree of spread, histology, socioeconomic status as predictors

<u>Definition of Model:</u>				
Cohort	All persons with a diagnosis of the ureter or renal pelvis who survived at least 1 day and were diagnosed using method 6 (histological verification at CCR)			
Entry	Date of diagnosis of upper urinary tract TCC			
Event	Diagnosis of bladder cancer occurring >=3 months post first cancer diagnosis			
Censoring	Diagnosis of non-bladder cancer, death or 31 st Dec 2007 whichever occurred first			
Variable	Categories	Hazard Ratio	(95%CI)	P(Wald)
Site	Ureter	1.00		0.03
	Renal Pelvis	0.66	(0.46, 0.96)	
Sex	M	1.00		0.84
	F	1.02	(0.68, 1.37)	
Year of diagnosis (ydg)	1986-1989	1.00		0.27
	1990-1994	1.32	(0.75, 2.32)	
	1995-1998	1.69	(0.96, 2.97)	
	1999-2003	1.26	(0.71, 2.24)	
	2004-2007	0.92	(0.44, 1.89)	
Degree of spread at diagnosis (stage)	Localised	1.00		0.76
	Regionalised	0.94	(0.63, 1.40)	
	Distant	0.81	(0.25, 2.61)	
	Unknown	1.25	(0.75, 2.09)	
Histology (hist)	TCC	1.00		0.01
	Papillary TCC	1.60	(1.12, 2.29)	
Age (age)	<65yrs	1.00		0.01
	65-74yrs	1.95	(1.25, 3.05)	
	75+ yrs	1.85	(1.12, 3.04)	
IRSAD quintile	Lowest	1.00		0.30
	Second	1.74	(0.90, 3.37)	
	Third	1.36	(0.73, 2.54)	
	Fourth	1.26	(0.66, 2.41)	
	Highest	1.00	(0.50, 1.98)	

Table 5: Cox Proportional Hazards Regression – Reduced Model: Time from upper urinary tract TCC diagnosis to diagnosis of bladder cancer as a second cancer– age, site of first primary, degree of spread, histology

<u>Definition of Model:</u>				
Cohort	All persons with a diagnosis of the ureter or renal pelvis who survived at least 1 day and were diagnosed using method 6 (histological verification at CCR)			
Entry	Date of diagnosis of upper urinary tract TCC			
Event	Diagnosis of bladder cancer occurring ≥ 3 months post first cancer diagnosis			
Censoring	Diagnosis of non-bladder cancer, death or 31 st Dec 2007 whichever occurred first			
Variable	Categories	Hazard Ratio	(95%CI)	P(Wald)
Site	Ureter	1.00		0.02
	Renal Pelvis	0.66	(0.46, 0.94)	
Histology (hist)	TCC	1.00		<0.01
	Papillary TCC	1.59	(1.12, 2.24)	
Age (age)	<65yrs	1.00		0.01
	65-74yrs	1.91	(1.24, 2.95)	
	75+ yrs	1.68	(1.04, 2.69)	

Discussion

The present study examined the elevated risk of experiencing a second invasive cancer diagnosis within specific sites of the urinary tract following an initial invasive urinary cancer. It also explored several potential factors that may elevate risk in the specific example of bladder cancer following cancer of the upper urinary tract.

There was a clear elevation of risk of bladder cancer following cancers of the renal pelvis and ureter in both males and females. This contrasted with less clear findings for bladder cancer following cancer of the renal parenchyma (with significantly elevated risk for females but not males) and of the urethra (with elevated risk for males, but indeterminate risk for females due to no observed second cancers). The differences between males and females in terms of standardised incidence ratios for second urinary cancers, reflects an underlying difference in incidence patterns between cancers of the upper urinary tract and cancers of the bladder, with most bladder cancers in NSW diagnosed in men, but a more even distribution of renal pelvis and ureter cancers between the sexes. This suggests a difference in causal factors between the two sites. McCredie *et al.* postulated that this could reflect underlying differences between the sexes in exposure to phenacetin containing analgesics which were removed from the market in the 1980's. (1).

When examining the reverse relationships, risks of renal parenchyma, renal pelvis, ureter and urethra cancers were all elevated following initial bladder cancer diagnosis for both males and

females. However, for all cancer sites that comprise the urothelium (including the renal pelvis, ureter, bladder and urethra), risks were highest when the second cancer was 'downstream' of the first. For example, risk of bladder cancer following renal pelvis was significantly higher than renal pelvis following bladder and risk of urethra cancers following bladder was significantly higher than bladder following urethra. However, this relationship did not hold for cancers of the renal parenchyma.

At the time of reporting, McCredie *et al.* did not examine the increased risk of renal cancers following invasive bladder cancer. It would be expected that field cancerisation caused by exposure to a common etiological factor would increase the risk of cancer across the entire urothelium. This theory does not propose a reason why cancer might develop more rapidly in one area of the urothelium compared to another, suggesting an equally elevated risk of renal pelvis cancer following invasive bladder cancer compared to the reverse relationship.

Alternatively intraluminal seeding suggests a mechanical action of spread by which cancer cells move within the urinary tract and 'seed' to another area. It would be expected that the risk of cancers occurring 'downstream' of a first cancer would be greater than the risk of second cancers in 'upstream' sites due to the directional flow of urine.

This pattern of excess risk supports the theory that second cancers of the urothelial tract may manifest due to a 'seeding' of cells from a prior cancer, assisted by the flow of urine. The influence of urinary flow in transporting cancerous cells is also supported by findings that patients who experienced urinary 'reflux' from the bladder to the ureter were more likely to develop cancers in the upper urinary tract (5). However, the significantly elevated risks for second cancers that were located upstream of the first also suggests that other mechanisms such as field cancerisation are important.

Several studies have examined risk factors for the occurrence of second cancers in the urinary tract, but most have been retrospective examinations of case series or small cohort studies within single institutions. Commonly cited factors predictive of bladder tumours following cancer of the upper urinary tract include tumour grade, multifocality, location (ureteric tumours having higher risk of recurrence than renal pelvic) and surgical procedure, with other factors such as sex and tumour size reported only on occasion (6) (7) (8) (9). Fewer studies have examined the risk of upper urinary tract cancers following bladder cancer, although one population-based analysis using data from the SEER database in the US found that tumour grade, stage and location were predictive of upper urinary tract recurrence (10).

As well as site of initial cancer, both the examination of SIRs and survival modelling demonstrated initial cancer histology as an important predictor of second cancer occurrence.

Those diagnosed with an initial papillary transitional cell carcinoma were more likely to develop a second bladder cancer than those diagnosed with a transitional cell carcinoma not otherwise specified. This is the first study that the author could locate that has examined histological sub-type as a predictor of a second urinary tract cancer.

Risk factors for bladder cancer in the general population would also be expected to be risk factors for second bladder cancers. Age at diagnosis was a positive predictor of second bladder cancer occurrence after controlling for other factors. As urothelial cancer incidence is rare under the age of 60, it is unsurprising that the risk of second bladder cancers increases with age. However, sex is also generally a risk factor for bladder cancer, with males more likely to be diagnosed than females, but sex was not significant in predicting second bladder cancers following cancers of the transitional cell carcinomas of the upper urinary tract. Factors such as smoking rates and to a lesser extent workplace exposure to aromatic amines are thought to increase the risk of bladder cancer for males, and are also risks for urothelial cancers in general. It seems likely that those diagnosed with an upper urinary tract cancer had a more equal distribution between the sexes of known bladder cancer risk factors than the population in general. Due to the high level of censoring due to death in some sub-groups in this study (such as the elderly) it is may be beneficial to re-examine the relationship of covariates using a competing risk model which will allow a more sophisticated inclusion of death within the model.

There are limitations in the registration practices of urinary cancers that need to be considered in the context of this study. Errors in the identification and coding of invasive versus non-invasive bladder cancers have been previously documented including within the SEER program and the NSW Central Cancer Registry (11) (12). Due to the variability over time in the coding of invasive versus in situ cancers, the SEER program has routinely reported incidence inclusive of both invasive and in situ to avoid what would otherwise result in an artefactual decrease for invasive cancers and increase for *in situ* cancers over time as coding practices were improved (11). Within NSW the reporting of in situ cancers is not mandatory, but following internal review of notifications many cancers have been recoded from invasive to in situ. A review in 2008 suggested that as many as 30% of transitional cell carcinomas and 70% of papillary transitional cell carcinomas were no longer classified as invasive following pathological verification (12). From 2006 onwards, bladder cancer registrations within the NSW CCR data were made more consistent, with all localised or unknown cancer registrations histologically verified.

In situ cancers and non-invasive low grade papillary carcinomas have been speculated to follow separate mutational pathways but both potentially leading to invasive cancers (2). As reporting of in situ urothelial cancers to the NSW Cancer registry is not compulsory, it is difficult to include in situ cases in any analyses due to the unknown coverage of this type of cancer. The present study was limited to invasive cancers which were histologically verified within the NSW Central Cancer Registry. It should be noted that this will have excluded a proportion of potentially eligible cases. Exclusion of both non-invasive cancers and non-verified invasive cancers is likely to reduce the power of any analysis, but is only likely to bias the results if exclusion of these cancers is related to the location of the tumour or the other covariates of interest.

Previous studies have found cancer stage to be a significant predictor of urothelial cancer recurrence with Carcinoma in Situ (CIS) to be associated with more frequent occurrence of subsequent cancers (8). The present study excluded all CIS and therefore had no scope to examine this relationship. When examining the order of cancer occurrence between different sites in the urothelial tract, it is important to recognize that a first cancer may have been preceded by a non-invasive cancer in another location. This may bias the results if non-invasive cancers are related to site of cancer. Additionally non-verified registrations may be linked to other factors such as year of diagnosis and the exclusion of these cases may interfere with the ability to examine this as a predictor of second cancer occurrence.

The elevated risk of bladder cancer following cancer of the renal pelvis and ureter clearly reflects the multifocal nature of urothelial tumors, and in 2004 it was proposed that rules for reporting 'second primary cancers' from cancer registries be adjusted to count the entire urothelium including the renal pelvis, ureter, bladder and urethra as one organ site (13). Thus, a second cancer occurring in any of these organs of the same histological group would be counted as a multiple cancer rather than as a second primary cancer and therefore excluded from incidence and mortality statistics. Previously, the renal pelvis and ureter were included with the renal parenchyma as one organ site and bladder was treated as a separate organ site.

Further analysis using registry data from other jurisdictions such as the SEER database in the USA is warranted to validate the findings here and extend the study to include non-invasive cancers. This study also underlines the complex and recurring nature of cancer diagnoses in the urothelial tract and the importance of accurate histological coding in cancer registries.

References

1. McCredie, M., Macfarlane, G.J., Stewart, J., Coates, M. *Second primary cancers following cancers of the kidney and prostate in New South Wales (Australia), 1972-1991.*, Cancer Causes and Control, 1996; 7: 337-344.
2. Kakizoe, T. *Development and progression of urothelial carcinoma.* Cancer Science, 2006;97:821-828.
3. Slaughter DP, Southwick HW, Smejkal W. *Field Cancerization in oral stratified squamous epithelium; clinical implications of multicentric origin.*, Cancer, 1953;6:963-8.
4. Hafner, C., Knuechel, R., Zanardo, L., Dietmaier, W., Blaszyk, H., Cheville, J., Hofstaeder, F., Hartmann, A. *Evidence for oligoclonality and tumor spread by intraluminal seeding in multifocal urothelial carcinomas of the upper and lower urinary tract.* Oncogene, 2001; 20: 4910-4915.
5. De Torres Mateos, JA, Banus Gassol, JM, PALou Redorta, J, Morote Robles, J. *Vesicorenal reflux and upper urinary tract transitional cell carcinoma after transurethral resection of recurrent superficial bladder carcinoma.* Journal of Urology, 1987;138: 49-51.
6. Kirkali, Z., Tuzel, E. *Transitional cell carcinoma of the ureter and renal pelvis.* Critical Reviews in Oncology/Hematology, 2003;47:155-169.
7. Ziguener, R.E., Hutterer, G., Chromecki, T., Rehak, P., Langner, C. *Bladder tumour development after urothelial carcinoma of the upper urinary tract is related to primary tumour location.* British Journal of Urology, 2006; 98: 1181-1186.
8. Matsui, Y, Utsunomiya, N, Ichioka, K, Ueda, N, Yoshimura, K, Terai, A, Arai, Y. *Risk factors for subsequent development of bladder cancer after primary transitional cell carcinoma of the upper urinary tract.* Urology, 2005;65:279-283.
9. Hall, M.C., Womack, S., Sagalowsky, A.I., Carmody, T., Erickstad, M.D., Roehrborn, C.G. *Prognostic factors, recurrence and survival in transitional cell carcinoma of the upper urinary tract: a 30-year experience in 252 patients.* Urology, 1998;52: 594-601.
10. Wright, J.L., Hotaling, J, Porter, M.P. *Predictors of Upper Tract Urothelial Cell Carcinoma After Primary Bladder Cancer: A Population Based Analysis.*, The Journal of Urology, 2009;181; 1035-1039.
11. Lynch, CF, Platz, CE, Jones, MP, Gazzinga, JM. *Cancer Registry Problems in Classifying Invasive Bladder Cancer.* Journal of the National Cancer Institute, 1991;83: 429-433.
12. H. McElroy, M. Arcorace, C. Cooke-Yarborough, J. Luker. *Anomalies in Registering Bladder Cancer in NSW.* [Abstract]. Asia Pacific Journal of Clinical Oncology. 2008; 4 (Suppl. 2), Abstract # 544.
13. IARC. *International Rules for Multiple Primary Cancers.* Lyon : IARC, 2004.

Appendix: Overview of Statistical Analyses

A1 Data Management

Three datasets were utilised for this study:

- The NSW Central Cancer Registry (CCR) reporting dataset accessed via the Cancer Institute NSW which included all registrations of invasive cancers in NSW from 1972 to 2007, linked to death information from the national death index.
- ABS resident population estimates for NSW for the years 1972 to 2007, accessed via the NSW Health Outcomes Statistical Toolkit (HOIST).
- Indices of Relative Socioeconomic Advantage and Disadvantage (IRSAD) for NSW 2006 (by postcode), accessed via the ABS.

The NSW CCR reporting dataset was made accessible through my secondment to the Monitoring, Evaluation and Research Unit at the Cancer Institute NSW for the purposes of the project.

The Estimated Resident Population for NSW by sex and single year of age (grouped for 85+) was used in the first analyses to calculate expected numbers of second cancers. The IRSAD indices were linked to the CCR dataset based on postcode of residence at time of diagnosis.

For both parts of this project the CCR dataset required transforming to link first and second cancer diagnoses for an individual based on the order of date of diagnosis. Third and later cancers were excluded for the purposes of this project.

The CCR variables used included:

Person level characteristics

- Registration number (to enable linking of first and second cancers)
- Sex (covariate of interest)
- Month of death (to allow for censoring at time of death, day was set at “15”)
- Year of death (to allow for censoring at time of death)

First cancer diagnosis characteristics

- Age at diagnosis (covariate of interest)
- Month of diagnosis (to allow calculation of date of entry to a cohort. Day was set at the “15th” of the month)

- Year of diagnosis (date of entry) (covariate of interest)
- Site of tumour (ICD-10 coding at a four digit level) (covariate of interest)
- Histology of cancer (ICD-03 coding at a four digit level) (covariate of interest)
- Method of diagnosis (to allow selection of cases where histology sighted at CCR (Method=6))
- Stage at diagnosis (highest degree of spread of cancer within 4 months of diagnosis (1=localised; 2=regionalised; 3=distant; 9=unknown)) (covariate of interest)
- Postcode of residence at time of diagnosis (to allow linkage to IRSAD index) (covariate of interest)

Characteristics of second cancer diagnosis

- Month of diagnosis (to allow calculation of date of event. Day was set at “15th” of each month)
- Year of diagnosis (date of event)
- Tumour site (to allow selection of events of interest)
- Tumour histology (to allow selection of events of interest)
- Method of diagnosis (to allow selection of events of interest)

A2: Evaluating excess risk of other urinary tract cancers following an initial urinary tract cancer diagnosis

Aim of the analysis:

Primary: To explore estimates of the elevated risk of second urinary cancers of different organ sites following an initial urinary cancer diagnosis.

Secondary: To explore patterns of excess risk based on site of first cancer, histology of first cancer and sex

Rationale for the approach taken:

Selection and definition of cancer sites

Distinct urinary organs were split into non-bladder and bladder for pragmatic reasons. Non-bladder urinary cancers included: Renal Parenchyma (main body of the kidney); Renal Pelvis (neck of the kidney that adjoins the Ureter); Ureter (joining the renal pelvis to the bladder); and Urethra. Due to coding rules, second primary cancers of the same organ site are excluded

from the CCR reporting database. In the case of urinary cancers, non-bladder cancers have historically been classed as the same organ site and bladder cancers as a separate organ site. This classification has changed since 2004 to include Renal Pelvis, Ureter and Urethra with Bladder and Renal Parenchyma as a separate organ site. Coding rules within the CCR have remained the same with reporting allowing for the change in rules. These historic rules allow examination of relationships between second primary bladder cancers following primary non-bladder cancers and vice versa, but do not allow investigation of multiple primary cancers of different sites within the non-bladder cancer group.

Calculation of Standardised Incidence Ratios (SIRs)

SIRs were calculated to allow investigation of rates of cancers within the population of interest (those diagnosed with an initial primary cancer of specific type) compared to the 'healthy' population (those not diagnosed with an initial primary cancer). The retrospective cohort approach is a commonly used epidemiological approach to calculate the number of observed versus expected cases required in the calculation of SIRs. This approach has also been the focus of previous work within the Cancer Institute NSW to develop a standard SAS program that can be adapted to explore second cancer occurrence for cancer types of interest.

Existing SAS program to calculate SIRs:

The existing SAS program allowed flexible investigation of different combinations of cancers. This program operates by the following steps:

1. CCR data is read in (using the re-shaped dataset with only first and second diagnosis and using all diagnoses from 1972 to 2007)
2. NSW population counts by age, sex and calendar year are downloaded for the period 1972-2007
3. Three cohorts are created based on the specification of the initial cancer of interest and sex:
 - Cohort C: >2 months post entry date to study end date;
 - Cohort B: >2 to 119 months; and
 - Cohort A: >2 to 59 months

Where study entry date is the date of diagnosis of a person's first cancer diagnosis. A person enters the cohort on this date and contributes person years of observation until diagnosis of a second cancer, death or study end (31 December 2007).

The construction of three separate cohorts allows calculation of SIRs for three distinct time periods, 3mo-<5yrs (cohort A), 5yrs-<10yrs (Cohort B-Cohort A) and 10 yrs plus (Cohort C-Cohort B) as well as an overall SIR (Cohort C). Diagnoses of second cancers made within the first three months were excluded to allow for multiple 'synchronous' diagnoses in which the order of occurrence may be uncertain.

4. Observed cases of each cancer type within each cohort are counted for each stratum (stratified by age last birthday, sex and calendar year of diagnosis).
5. Population incidence of each cancer type is calculated within each cohort for each stratum
6. Person years of observation are calculated within each cohort for each stratum
7. Expected cases of each cancer type are calculated within each cohort for each stratum:

$$\text{Expected}_{\text{stratum}} = \text{Incidence}_{\text{population}} * \frac{\text{PYO}_{\text{stratum}}}{\text{PYO}_{\text{population}}}$$

8. SIRs are calculated in each cohort:

$$\text{SIR}_{\text{cohort}} = \frac{\sum_{\text{strata}} \text{Observed}_{\text{cohort}}}{\sum_{\text{strata}} \text{Expected}_{\text{cohort}}}$$

9. 95% confidence intervals are constructed based on the Poisson distribution.

Adapting the existing SAS program:

The existing SAS program could be used to calculate SIRs for each of 54 programmed second cancer types for the specified initial cancer type and results could be output either individually for each sex or combined. Slight modification of the program was required to disaggregate the grouping of "Kidney" for definition of second cancer type which initially included all non-bladder urinary sites of Renal Parenchyma, Renal Pelvis, Ureter, Urethra and other urinary organs not otherwise specified. These were re-defined as separate cancer types.

Additional factors such as histology were also included in the definition of initial cancer type to allow more specific investigations to be carried out.

Potential limitations of this approach

One of the main potential limitations of the SAS algorithm used is the issue of interpreting multiple comparisons. With a large number of combinations of cancers able to be investigated

and then also broken down into sub-groups such as by sex and histology, it would be expected that a number of 'significant' findings in such a large number of comparisons could occur by chance alone. However, in the present study the number of analyses were limited and based on *a priori* hypotheses which reduce the significance of this issue.

A3: Determining predictors of second cancer occurrence through Cox Proportional Hazards regression modelling

Aim of the analysis:

To determine what factors predict the occurrence of an invasive bladder cancer following an upper urinary tract cancer.

Rationale for the approach taken:

Survival analysis was considered an appropriate technique for examining predictors of second cancer occurrence. The main reason for this was that the ability to observe the occurrence of a second cancer could be highly biased due to censoring when a person dies. As cancers can have quite high death rates, and most covariates of interest could potentially be related to death rates, the impact of censoring could be significant. It was possible for one person to be diagnosed with many cancers. So as to allow examination of the relationship between just two cancers of interest, a person was considered censored at the date of a non-bladder cancer. Survival analysis provides more information on the rate of cancer occurrence over time that could not be gained by just examining SIRs.

The potential issue with this approach given the dataset included the possibility of missing data due to loss of follow-up when people move interstate or overseas. However, due to the co-operation between jurisdictions within Australia and the efforts made in matching death records, this issue is likely to have minimal effect.

Another issue is that a high proportion of cases of first cancers in older age groups will have died before having the chance to acquire the second cancer. As there are differences in death rates at different levels of covariate, it may be misleading to regard deaths as the same as censored observations. Accordingly, proportional hazards regression incorporating competing causes/risk is likely to be a more appropriate approach. It is the intention that this method of analysis will be undertaken over the coming months to increase to validity of the findings before further considering this paper for publication.

SAS Version 9.2 was used for all analyses. SAS is the standard statistical analysis software supported within the Cancer Institute NSW and also allows for analysis of large datasets.

Overview of the analysis steps taken:

a) Assessing the distribution of each variable

Each variable was inspected for outliers and distribution was assessed.

There were 1700 observations with a valid time to event (>0.0yrs). Time to event was highly right skewed which is not unusual for this type of data, but the nature of the administrative dataset allows a very long follow-up period resulting in a large range of values from 0.04-21.8 years with a median follow-up time of 1.8 years.

Two covariates could potentially be included in the analysis as continuous variables: age and year of diagnosis (ydg). Histograms for these two variables are presented in Figure A1. Year of diagnosis, as expected had a fairly flat distribution and as there was no prior reason to assume that this variable would be linearly related to the occurrence of bladder cancer, it was treated instead as a categorical variable labelled “period of diagnosis” with 4-5 yearly groupings. Age appeared to be reasonably normally distributed with a slight tail to the left. Based solely on the distribution, it did not appear necessary to transform this variable and it was included in the continuous form for further investigation of the survivorship function by different levels of the covariates.

An overview of the proportional distribution of categorical variables across groups was presented in table 3 on page 18. No significant issues were identified.

b) Assessing the overall survivorship function

The Kaplan-Meier survival curve was calculated (Figure A2). This showed that the overall proportion of the cohort experiencing the event of interest (diagnosis of a bladder cancer) was quite low and that the rate of second cancers was fastest in the first 2.5 years following initial cancer.

c) Assessing the survivorship function by each level of each covariate

Kaplan-Meier survival curves were examined for each level of covariate to examine potential predictors and log(-log) survivor functions were plotted against log survival time to help identify any potential issues with assuming proportionality of hazards (Figures A3). The only covariates that appeared to be clearly related to time to second cancer at the univariate level were age, site of first cancer and histology. A log-rank test confirmed these relationships to be significant (Table 3 on page 18). Year of Diagnosis (grouped into five periods) appeared not to be strongly related to second cancer occurrence and there was no uniform increasing or decreasing trend across periods. The lack of trend was supported by a non-significant log-rank test for trend ($p=0.51$). Age did appear to be related to time to second cancer with a significant log-rank test for trend ($p=0.03$). However, the trend did not appear linear with the oldest two age groups appearing to have similar survivorship. There was also an indication that hazards may not be proportional at all survival times for age. Based on these findings, and to aid in simplicity of interpretation, the decision was made to include age as a categorical variable and test the proportional hazards assumption further at a multivariate level.

d) Building the Cox Proportional Hazards Model

All potential covariates were initially included in the model. This “full” main effects model showed a significant effect of site, histology and age group on time to second cancer. All other covariates were non-significant at the 5% level. As the main purpose of the analysis was testing the significance of potential predictors, the full model was of interest and the results reported.

However, a more parsimonious main effects model was also investigated. Covariates were removed from the full model one by one based on descending p values until only significant predictors remained. No further covariates became significant and the reduced model remained with three covariates site, histology and age group. There are some limitations in constructing a regression model in this way. For example, confidence intervals for effects may be overly narrow. However, both the full and reduced model are reported and effects for the three significant predictors remained reasonably consistent.

Interaction effects were also investigated within this reduced model. There was no prior clinical rationale for examining particular interaction effects. However, it was possible that the effect of histology could vary by site and also that the effects of both histology and site could vary by age. Five interaction terms were constructed by creating dummy variables (Age_65-74*HistTCC, Age_65-74*SiteRenalPelvis, Age_75+*HistTCC, Age_75+*SiteRenalPelvis, HistTCC

*Site_RenalPelvis) and were included individually one by one in the reduced main effects model. None were significant at the 5% level and so were not included in the model.

e) Model Diagnostics

(i) Assessing the Proportional Hazards assumption

Schoenfeld residuals were used to assess the assumption of proportional hazards for each covariate in the reduced main effects model. The advantage of examining Schoenfeld residuals is that it allows examination of which covariates may be violating the proportional hazards assumption. Residuals were plotted against survival time (Figure A4). There were some slight deviations at low and high survival times for all covariates. To further investigate this issue, time-dependent versions of the covariates were included in the model by including covariate by time interaction terms. None of the time-dependent interaction terms were significant at the 5% level and the overall test of non-proportionality was non-significant (Wald $\chi^2=3.25$, $p=0.52$). This provides support for the appropriateness of the Cox model.

(ii) Assessing overall model fit and influential observations

Deviance and Martingale residuals were examined to assess overall model fit. Figure A5 shows the respective residuals plotted against the linear predictor. There was no overall skew towards high or low values of the linear predictor. However, there did appear to be a number of observations with very high positive residuals. Further examination of the deviance residuals plotted in conjunction with LMAX values (Figure A6- high LMAX values indicated by larger diameter circles) suggested that there were a number of observations with high influence and high positive residuals. Observations with residuals higher than 2.5 were inspected, but no obvious data errors existed. These were all observations that experienced the event of interest, had very short survival times and included a mix of covariate values. Removing the highest 5 values (with deviance >3) did not make any significant difference to covariate estimates.

Overall, the residuals appeared to reflect the divergence between censored observations which tended to have longer survival times than predicted (negative residuals) and non-censored which had shorter survival times than predicted by the model (positive residuals). This is consistent with a population that experiences a low event rate with most events occurring in the first few years but with potentially very long follow-up period for censored observations due to the passive nature of surveillance (determined by matching to death records rather than active participation).

The model explains only a small proportion of variation in time to event. However, there appears to be no overall bias for particular covariate values and towards particular survival times. Given the main purpose of the analysis was to examine potential covariate predictors rather than predict survival times, the model appears appropriate for this purpose.

f) Final Model

Following assessment of model diagnostics, the reduced main effects model was accepted as the most parsimonious model summarising the significant predictors of time to bladder cancer.

Figure A1: Histograms - Age at first cancer diagnosis and year of first cancer diagnosis

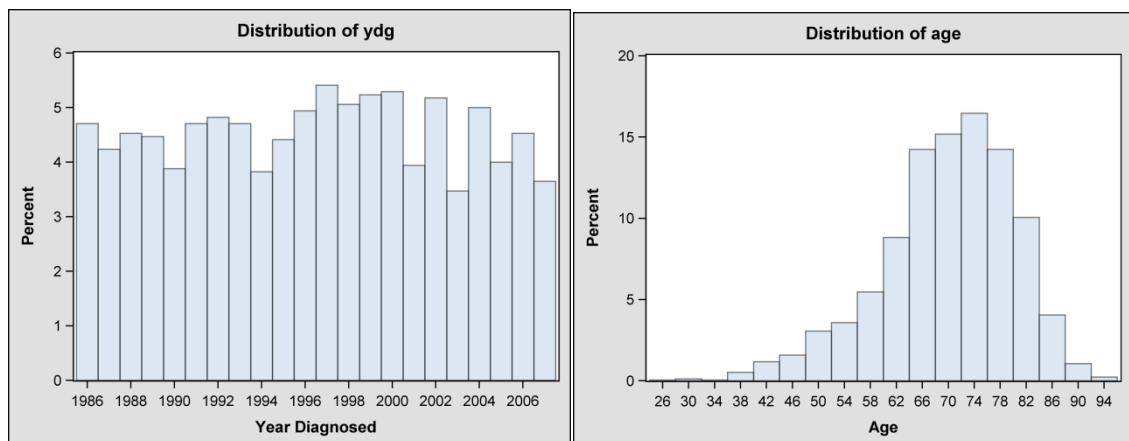


Figure A2: Kaplan-Meier survival curve for reduced main effects model

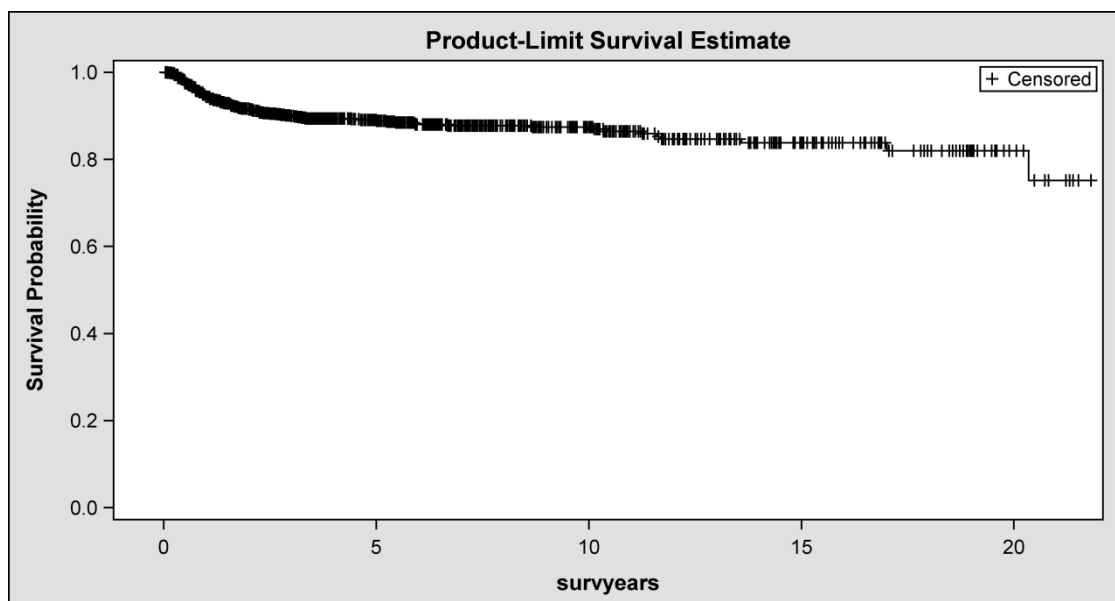
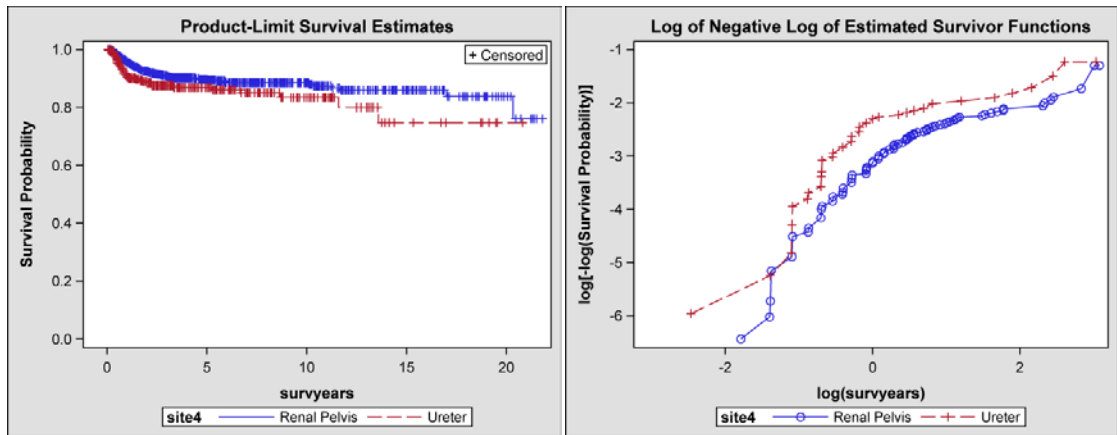
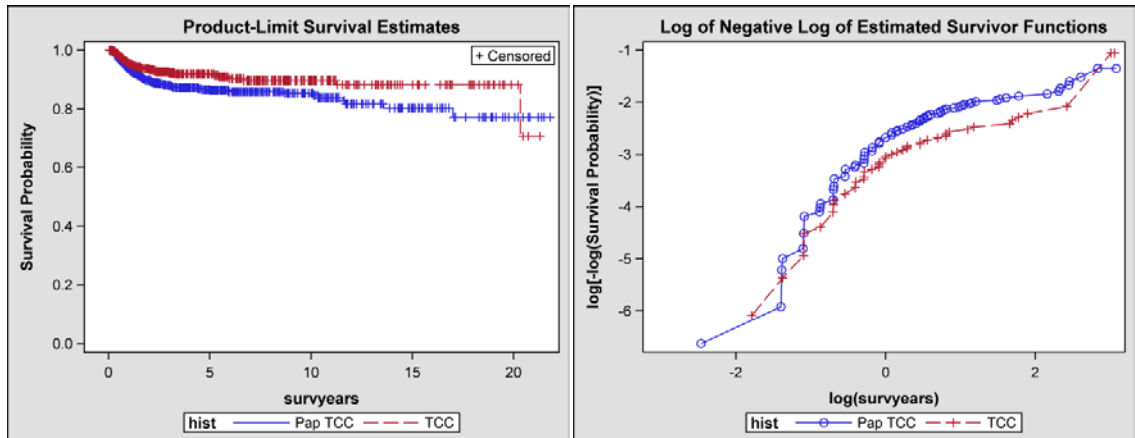


Figure A3: Kaplan-Meier survival curves and log(-log) survivor functions for each covariate

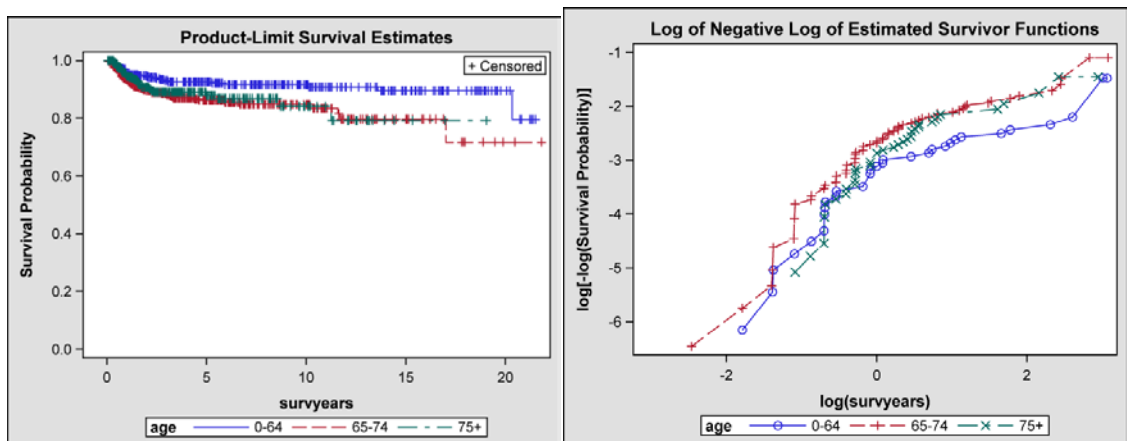
a) Site of First Cancer



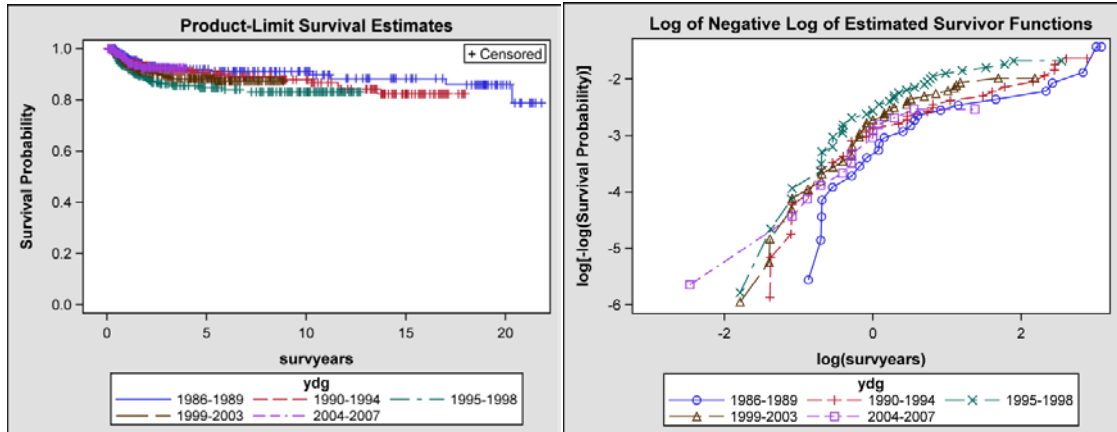
b) Histology



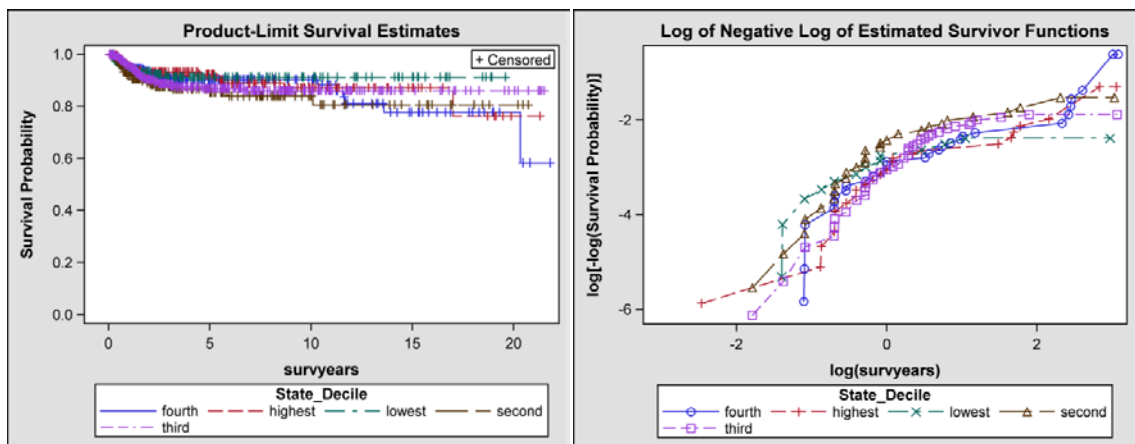
c) Age Group



d) Period of Diagnosis



e) Index of Relative Advantage and Disadvantage



f) Sex

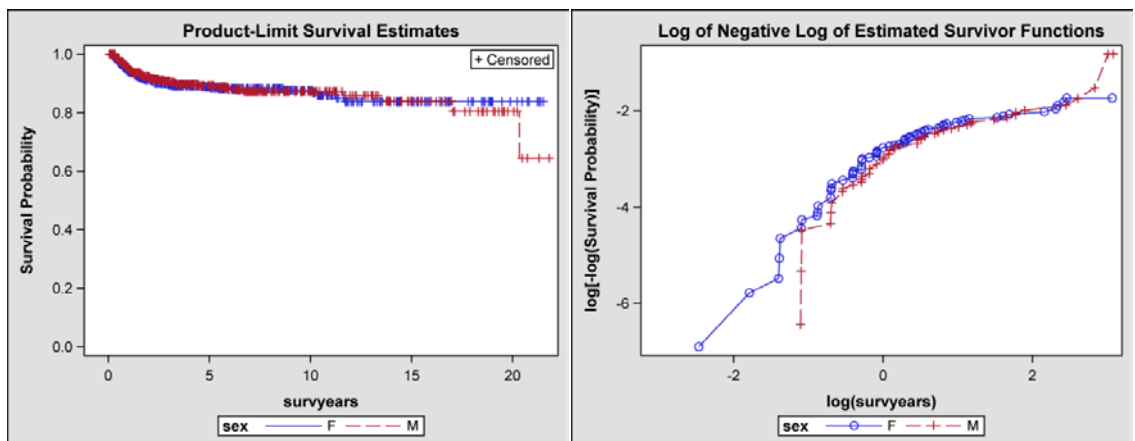


Figure A4: Schoenfeld Residuals for each predictor

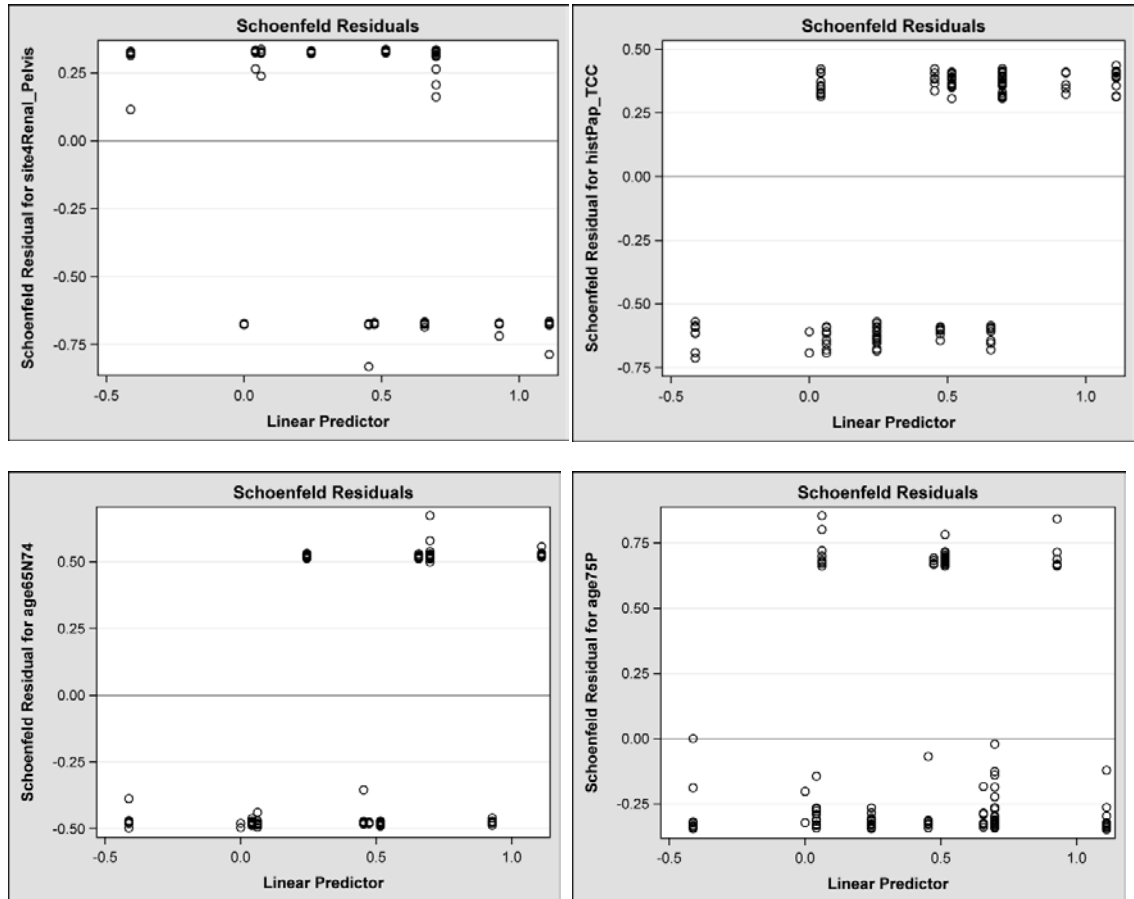


Figure A5: Martingale and Deviance Residuals for model

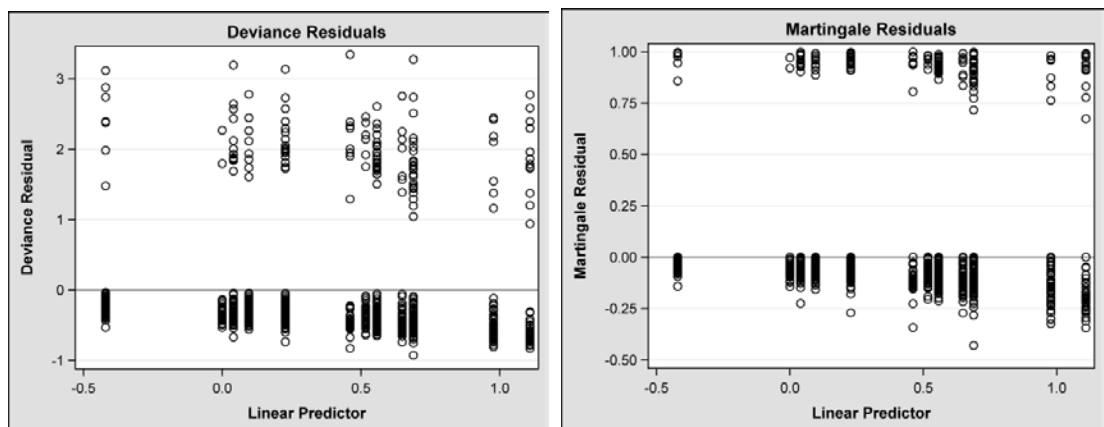
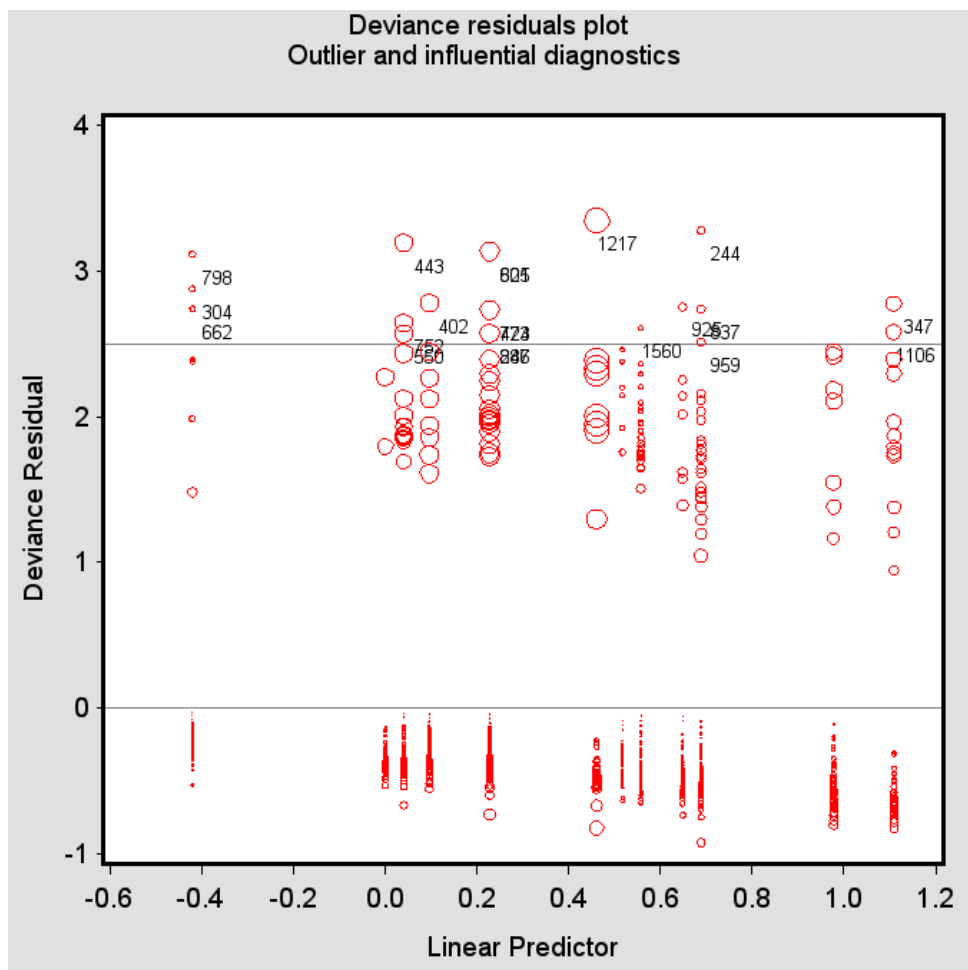




Figure A6: Deviance Residuals plotted with influence diagnostics (larger diameter circles indicating higher LMAX scores).



PART C: Project 2 – Multiple Imputation to address a data artefact for the degree-of-spread variable in the NSW CCR for the period 1993 – 1998: Lung Cancer as a test case

Location and Dates:	Cancer Institute NSW, Sydney, November 2010 – April 2011
Context:	This project was completed as part of the approved program of work within the Monitoring Evaluation and Research Unit at the Cancer Institute NSW. Heidi was seconded to MERU from another area of the CINSW to work on this project on a part-time basis. This project addressed a previously reported coding issue within the NSW CCR.
Contribution of student:	Heidi completed all design of study, data management, analyses, and write-up/presentation of results. Some extra input was required from the biostatistics team in terms of data extraction and creation of new variables.
Statistical issues involved:	<ul style="list-style-type: none"> • Logistic Regression Model building • Missing Data mechanisms • Multiple Imputation • Test validation • Kaplan-Meier Survival Curves
Declaration:	I declare that I have undertaken this project independently and have not submitted this work for previous academic credit.
Signed:	

Supervisors Name:	Dr Stephen Morrell (Co-supervisor with Ms Deborah Baker)
Statement:	Heidi has completed this piece of work diligently and in a timely fashion. She has performed all requested analyses and has shown initiative in undertaking different and further statistical analyses. Heidi has acquired a good understanding of the underlying issues involved in imputing missing values in cancer registration records. Her work will provide a sound basis for rectifying a major missing data problem in NSW cancer registrations.
Signed:	

Multiple Imputation to address a data artefact for the degree-of-spread variable in the NSW CCR for the period 1993 – 1998: Lung Cancer as a test case

Abstract

An artefact in degree-of-spread coding within the NSW Central Cancer Registry (CCR) data occurred for all solid-tumour cancers, excepting breast and melanoma, diagnosed for the period 1993-1998 (1). This resulted in a decrease of cases coded as localised and an increase in cases coded as unknown. The cause was the introduction of the Electronic Notification System (ENS). Cancers with regionalised or distant degree of spread were unaffected. This artefact has implications for using the degree of spread variable within the CCR and imposes limitations on analyses. This paper outlines the scope of the problem for one cancer type – lung cancer – and investigates multiple imputation (MI) as a method for addressing the problem. Cases with “unknown” degree of spread that included electronic notifications were classified as having missing values for the degree of spread variable for the period 1993-1998. Cases were then re-allocated to the localised and unknown categories based on MI, using a logistic regression model as the basis for prediction. The model produced plausible results that appeared to correct the artefact and were consistent across sub-groups. Independent validation in a distinct time period suggested that the model had reasonable prediction accuracy (69%) for coding localised cases. Survival was significantly poorer for localised cases within the period 1993-1998 based on imputed data compared to original coding, but imputation had no effect on survival when degree of spread was unknown. The MI model tested was specific for lung cancer but could also be modified and tested on other cancer types affected by the data artefact.

Contents

Section 1: An overview of the problem	40
Section 2: Building a predictive multivariate model for localised versus unknown degree of spread for lung cancer	49
Section 3: Multiple Imputation to correct the degree-of-spread data artefact for lung cancer cases.....	56
Section 4: Impact of correcting the data artefact on survival estimates.....	64
Section 5: Discussion and Conclusions.....	69
References	71
Appendix A: Summary of potential predictors for degree of spread	72
Appendix B: Logistic Regression results for full model (model 1)	84

Section 1: An overview of the problem

Background: The NSW Central Cancer Registry

The NSW Central Cancer Registry (CCR) receives notifications of all malignant cancers diagnosed in NSW. The CCR is managed by the Cancer Institute NSW for the NSW Department of Health (NSW Health), and operates under the authority of the Public Health Act of 1991. The Registry maintains a record of all malignant cancer cases diagnosed in NSW residents since 1972. However, notification of malignant neoplasms has been a statutory requirement for all notifying institutions in NSW since 1986.

These institutions include public and private hospitals, departments of radiation oncology, nursing homes, pathology laboratories, outpatient departments and day procedure centres. When any of these institutions diagnose or treat someone with malignant cancer, they are required by law to notify the NSW Central Cancer Registry. Notifications of cancer in NSW residents are also received from cancer registries in other states and territories.

The NSW Central Cancer Registry aims to monitor the number of new cases of cancer and deaths from cancer in NSW and assist in cancer prevention and control by producing descriptive analyses of cancer incidence and mortality trends, facilitating epidemiological and clinical research, and supporting planning, evaluation and monitoring of services and screening programs.

To this end, the NSW CCR reporting database contains a mixture of demographic and clinical variables. It records the year and month of birth, death and diagnosis of each cancer case in NSW, plus basic demographic variables such as sex, Aboriginal and Torres Strait Islander status and postcode. Clinical variables such as the site of cancer, histology of cancer, method of diagnosis and degree of spread are coded by medical coders within the NSW CCR based on information provided within notification reports.

Degree of spread in the NSW CCR is a summary measure based on cancer staging at first presentation. It is derived by the CCR from the maximum extent of disease based on all reports and notifications dated within four months of the date of diagnosis. Degree of spread reported here follows the international coding guidelines for summary stage adopted by several international groups including the World Health Organization and the International Association of Cancer Registries (2).

Degree of spread is grouped as:

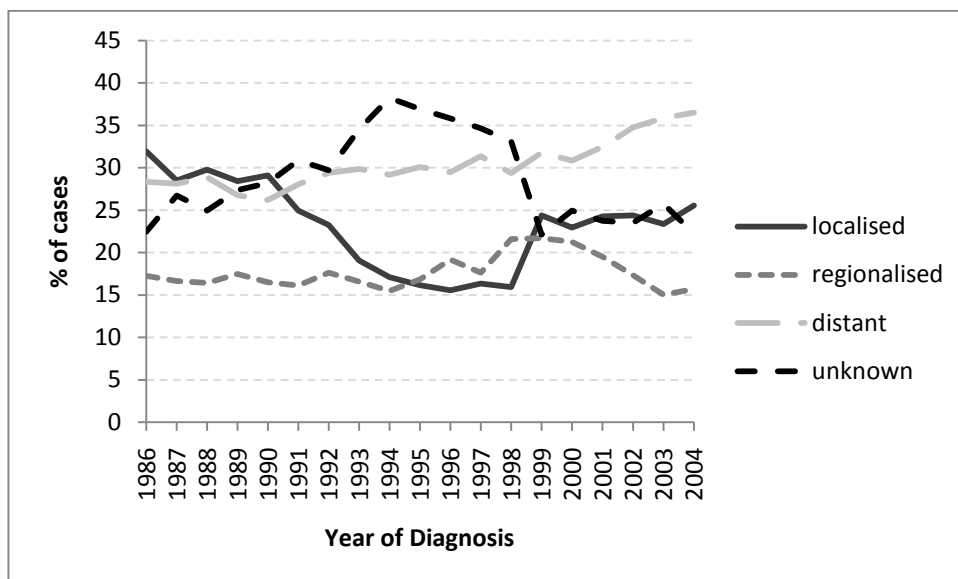
- (i) localised (assumed to predominantly consist of TNM Stage 1 but may include some Stage 2)
- (ii) regionalised (assumed to be predominantly TNM Stage 2 and most of Stage 3)
- (iii) distant (predominantly TNM Stage 4 cancers).
- (iv) Some cancers are classified as unknown degree of spread for which staging information is inadequate or has not been collected.

Background: Data Artefact

Barraclough *et al.* described an artefact in NSW Central Cancer Registry (CCR) data that occurred for all solid-tumour cancers, excepting breast and melanoma, diagnosed for the period 1993-1998 (1). For these cancers within this period, the proportion of 'localised' cancer cases reported was approximately 5% lower than expected and was mirrored by an artefactual increase in 'unknown' degree-of-spread cases. This was caused by the introduction of the Electronic Notification System (ENS) which only affected the accuracy of coding of localised cancers, with regionalised and distant degree-of-spread cancers unaffected. This artefact has implications for using the degree of spread variable within the CCR and imposes limitations on analyses.

Figure 1 presents the proportion of lung cancer cases by degree of spread coding category for the years 1986-2004 based on year of diagnosis. There was a marked rise in unknown cases from 1992 to 1993, mirrored by a drop in localised cases. Lung cancer shows an under-estimation of localised cases of almost 10 percentage points from what would have been expected during this period – much higher than the overall cancer artefact reported by Barraclough *et al.* These trends were caused by the absence of a 'localised' category within the ENS and were corrected in 1999 when the ENS was amended. This artefact can be even more clearly seen when considering just localised and unknown lung cancer cases (figure 2).

Figure 1: Percentage of Lung Cancer Cases by Degree of spread at diagnosis – 1986-2004



It is possible to ascertain for each diagnosed case, the method(s) of notification received by the CCR using the batch number of each notification episode. Multiple notifications may be received for a single case, for example from both a hospital and pathology lab. All notifications dated within 4 months of diagnosis are used to assess the degree of spread at diagnosis, and the case is categorised based on available information. The introduction of electronic notifications was a gradual process with almost 100% of lung cancer cases based on manual notifications (M) prior to 1993 which then dropped to about 60% by 1994, 45% by 1999 and 30% in 2004. This was mirrored by a gradual increase in cases notified by electronic means only (E) and a less gradual increase in cases notified by a mixture of manual and electronic means (EM) (Figure 3). There were a small number of cases diagnosed prior to 1993 that included electronic notifications. This is plausible given the possibility of a delay between an episode of care (eg. seeing a patient within a hospital) and a notification being sent. Additionally, as degree of spread is coded based on notifications within a four month window, cases diagnosed at the end of 1992 may have had notification episodes in 1993.

Figure 2: Percentage of Unknown and Localised Lung Cancer Cases by Degree of spread at diagnosis – 1986-2004

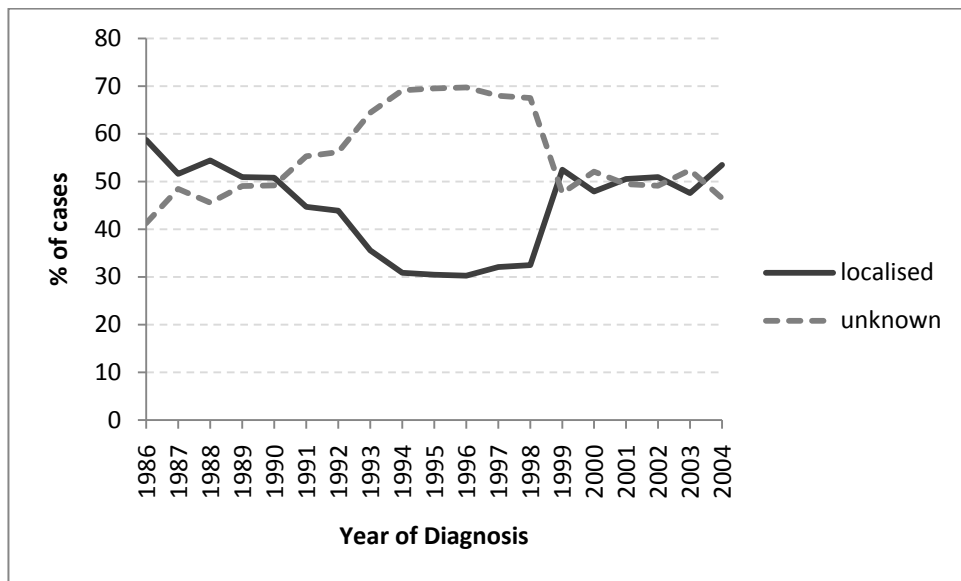
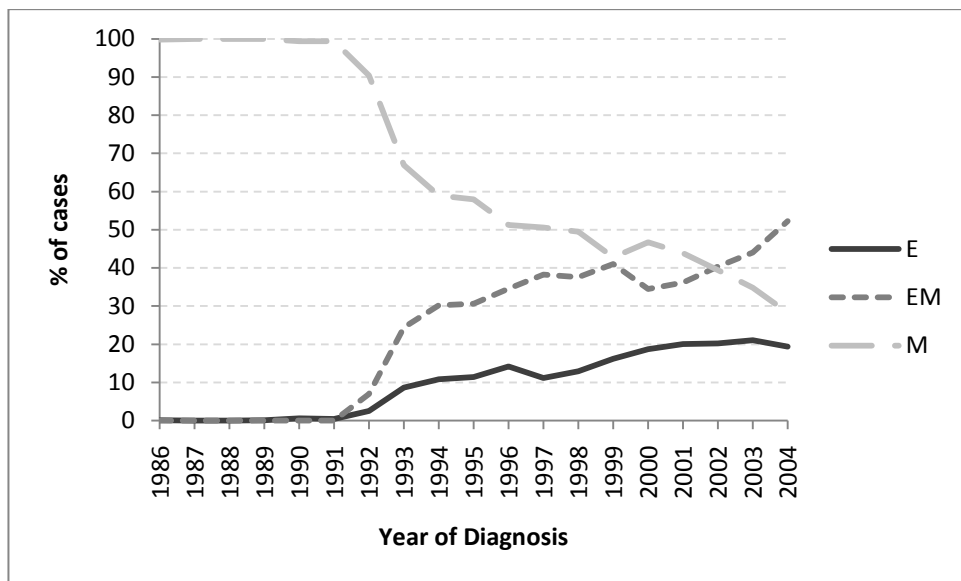


Figure 3: Percentage of Unknown and Localised Lung Cancer Cases, by Method of Notification, 1986-2004



Examination of trends in the ratio of localised versus unknown degree of spread by year of diagnosis (Figure 4), suggests that the known data artefact (evidenced by under reporting of ‘localised’ cancers and over-reporting of ‘unknown’ cancers) includes cases notified by electronic means only (E) and cases where some manual episodes were recorded in addition to the electronic episodes (EM). The cause of the artefact was the exclusion of a category for ‘localised’ cases on the ENS which meant that the notifying party could not select this response. For lung cancer, the information required for coding degree of spread was likely to have come from either: (i) imaging which would have been noted within hospital reports, or (ii)

pathology reports. Electronic notifications during this period would generally have comprised hospital reports, mostly from the public sector, with private hospitals commencing electronic notification at a much later time. Pathology reports are received manually which means that cases diagnosed based only on electronic notifications only would not have included pathology.

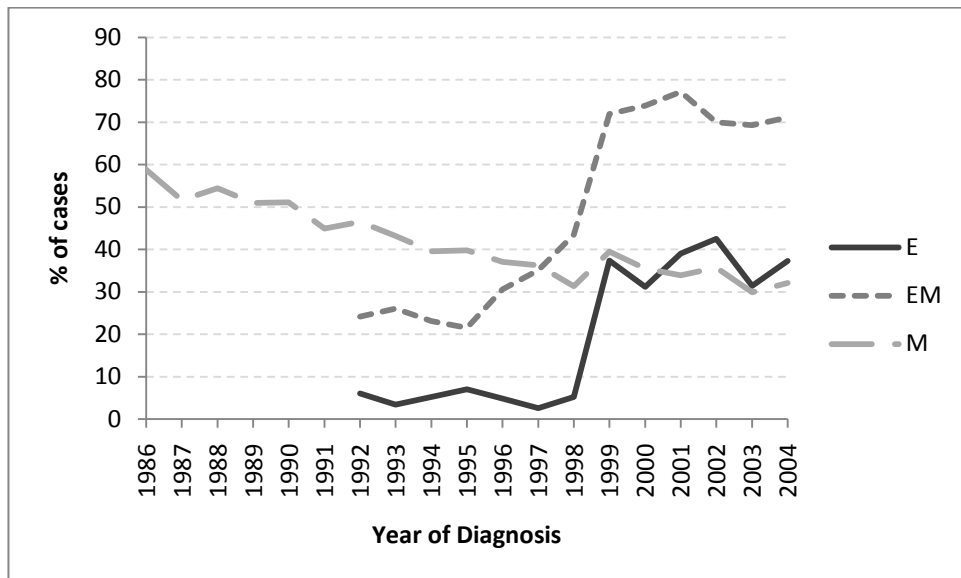
The implication of the changeover to ENS was that some information regarding degree of spread based on imaging may have been lost if the hospital report was submitted using the ENS. However, information gained from pathology reports or from hospital reports not submitted via the ENS was still available for coding this category. For the EM-notified cases this means that during 1993-1998 there may have been information available to code degree of spread if the manual notifications included pathology or private hospital reports, but that once the ENS was amended, the information gained from public hospital notifications would have been included again, increasing the amount of information available within the EM category. For cases notified by electronic means only (E) only a very small proportion of cases were coded as localised during this period. In a very small number of situations there may have been adequate information provided elsewhere on the electronic notification to code the case.

Figure 4 shows the clear and rapid increase in proportion localised between 1998 and 1999 for both the E and EM groups following the amendment of the ENS. From 1999 onwards, cases within the EM group show a much higher proportion of localised cases compared to either the E or EM groups. This can be explained by two factors. Firstly, these cases are more likely to have both imaging and pathology information which is likely to increase the chance of being able to accurately code degree of spread. The manual only group may have pathology but are less likely to have information from public hospitals and the electronic only group are less likely to have pathology. Secondly, the EM group were most likely to have a diagnosis made based on histopathology (including verified by CCR and unverified) rather than other means, compared to the E and M groups. Histopathology, particularly if it had been verified within the CCR, was a strong predictor of the ability to categorise cases as “localised” compared to “unknown” (data presented in section 2 and Appendix A).

Those cases notified by manual means only (M) show a steady decline in proportion localised over time, but do not show evidence of a data artefact for the period 1993-1998. The decline in proportion localised must have been driven by non-ENS related factors prior to 1993, but following 1993 may also be explained by a proportion of cases ‘shifting’ into the electronic plus manual category. Within this category it is more likely that multiple notification episodes

occurred and a clear link was found between number of notification episodes received and the ability to categorise the cancer as “localised” (data presented in section 2 and Appendix A).

Figure 4: Proportion of Localised cases (of all localised and unknown) by method of notification, 1986-2004



These figures support the finding that the data artefact is linked to the introduction of the ENS in 1993, with manual cases unaffected during the period 1993-1998. Additionally it appears that cases were affected even if notifications were received manually and electronically (EM) rather than just by electronic means only (E).

Missing data

The proposed approach to dealing with this data artefact relies on the assumption that we can justify classifying the problematic data as ‘missing’. Missing data are the result of a ‘missing value mechanism’, which may not be clearly identified. In the case of surveys, and administrative health information, data can be missing for 6 generic reasons:

- (i) the question may not have been asked of a respondent or patient;
- (ii) the question was asked but the subject did not respond for various reasons including refusal or ‘no comment’;
- (iii) the question was asked and the respondent did not know the answer but there was no ‘not known’ response category in the data collection instrument or database;

- (iv) the question was asked and the respondent responded but for some reason the response was not recorded;
- (v) the question was asked, the respondent answered, the response was recorded but the data were not entered;
- (vi) the data are missing legitimately, as in a 'not applicable' response, but this category may not be in the data collection instrument or database.

Missing data can cause problems with analyses by reducing statistical power and potentially introducing bias into estimates (3). If data are missing completely at random (MCAR), the cases with missing data are akin to a random sample of the observed cases. While this may reduce power, it is unlikely to bias estimates.

If data are not MCAR, then they are often classed as missing at random (MAR) or missing not at random (MNAR). Both are problematic as they will likely reduce power and bias estimates. MAR allows the probability of missing data for a variable X to be dependent on other variables in the dataset but not on X_i , where i is the value of X for an individual observation. In the case of MNAR the probability of the data being missing will be dependent on the value of X_i . MNAR data is difficult to address without further knowledge of the missing data, but MAR data can be addressed either through limiting analyses to subgroups related to the missing data pattern or via imputation processes. In practice it is often difficult to determine whether data are MAR or MNAR (4).

Missing data for the degree-of-spread variable 1993-1998

In normal circumstances, the 'unknown' category is a legitimate response for degree-of-spread and is coded by NSW CCR staff when the notifications received do not provide sufficient information to ascertain a degree of spread. The notifying party such as hospital or pathology lab may, given further testing/pathology, be able to determine degree of spread, but the NSW CCR coding staff, are not privy to decision-making regarding further investigation.

In the artefact period, there was an additional causal factor at play whereby the answer may have been known by the notifying party, but as there was no category for indicating degree of spread as 'localised' on the electronic notification system, this piece of information was not conveyed to the registry. This aligns most closely with cause (i) above – the question was not asked of the notifier, or cause (iv) the response was not able to be recorded.

In the period 1993-1998 all cases coded as 'unknown' based on electronic notifications (E or EM) are associated with an element of uncertainty. If the data collection problem had not existed then they may have been coded as localised rather than unknown. For E cases, all 'unknown' cases could be considered missing. For EM cases, as some information was still received manually, a proportion of cases were likely to have been affected by the data collection issue with an unknown proportion unaffected. However, as there is an element of doubt associated with these values we could therefore consider all cases notified as unknown by E or EM means to be 'missing' rather than 'unknown'. From this point on, these cases will be referred to as 'missing' and the aim will be to re-distribute the cases to the categories 'unknown' and 'localised' in a valid and informative manner.

In the current situation we know the data is not MCAR, as the probability of missing data is known to be dependent on both year of diagnosis and notification method. We have knowledge of the missing data mechanism based on the knowledge of the introduction of the ENS in 1993 and the amendment of the system in 1999. However, particularly for cases that received both electronic and manual notifications in this period, the mechanism for identifying when adequate information was available for coding is not able to be clearly identified based on available data. While we assume that the missing data is MAR, there is a possibility that it is MNAR. It should be noted however, that the missingness of the localised or unknown degree-of-spread category during the artefact period did not depend on whether the cancer's true degree of spread was localised or unknown. In the current study, imputed data patterns will be assessed to determine the plausibility of the MAR assumption. Additionally, the MI model will be applied to data in a distinct time period with known values for localised and unknown to test the sensitivity of the model in a situation where the MAR assumption holds and one where it does not. Sensitivity analysis has been proposed as a useful technique to assess the appropriateness of the MAR assumption (4).

Multiple Imputation to correct for missing data

There are various ways of dealing with missing data in analyses. The most common method is complete case analysis, whereby only cases without missing data are included. In the current situation, this is not feasible due to both the number of cases that would be designated as missing and the relationship between missing data and both time period and notification method. A solution to the problem proposed by Barraclough *et al.* was to consider using grouped data only (eg. group unknown and localised cases together for this period). However, this limits any analyses that aim to investigate the effect or outcomes of localised cancer.

Alternatively, analyses could be limited to a time period outside the known artefact period, such as using cases diagnosed from 1999 onwards. Again this limits the types of analyses able to be undertaken as often trends over long time periods are required to examine epidemiological relationships.

Imputation is a process by which the missing data are replaced with plausible values, often based on knowledge of other variables in the dataset. Single imputation can result in spuriously precise estimates. Multiple imputation takes into account the uncertainty introduced by estimating missing values. Generally only a small number of imputations (between 3 and 10) are required, and the inter-imputation variability can then be used to adjust the error component of subsequent analyses (5).

It is known that the degree of spread variable is related to other variables within the CCR database and is commonly used to monitor differences in survival patterns for most cancers (6). For this reason, it is likely that a logistic regression model will provide a reasonable level of prediction of localised versus unknown cases. The next section examines potential predictive variables and then potential logistic regression models that could be used within the multiple imputation process.

Section 2: Building a predictive multivariate model for localised versus unknown degree of spread for lung cancer

Assessing potential predictors of localised versus unknown degree of spread

A range of variables in the CCR database were investigated to assess their association with localised versus unknown degree of spread. The first aim was to identify variables that were significant predictors of localised versus unknown degree of spread. Associations were assessed for all localised and unknown degree-or-spread lung cancers diagnosed in years either side of the artefact period (1986-1992 and 1999-2004) and for all notification methods. Lung cancer cases were identified based on ICD10 codes C33-C34. The decision was made to exclude all cases diagnosed by death certificate only as very limited information is available for these cases and degree of spread is nearly always unknown. There were 25,082 cases of unknown or localised lung cancer diagnosed between 1986-2004, of which 16,922 fell in the two periods 1986-1992 and 1999-2004. Within these latter periods, 50.5% were localised and 49.5% were unknown.

Given the relationship of the missing data to two factors – diagnosis period and notification method – the second aim was to assess the consistency of variables as predictors of degree of spread. Firstly, association with degree of spread was assessed across three time periods (1986-1992 – period 1; 1993-1998—period 2; and 1999-2004—period 3) within manual notifications only, and secondly, across each notification method for time period three (1999-2004).

There were 16,467 cases diagnosed from 1986-2004 notified by manual means only, of which 44.4% were localised and 55.6% unknown. Of these cases, 53.3% were in period 1, 27.7% in period 2, and 19.0% in period 3. There were 8,002 cases diagnosed in period 3 across all notification methods with 50.5% localised and 49.5% unknown. Of these cases, 19.3% were based on electronic-only notifications, 41.6% on electronic and manual, and 39.1% on manual only.

The variables investigated fell into three categories: (i) Basic demographic variables (age, sex, socioeconomic status, area health service of residence, Aboriginal and Torres Strait Islander status); (ii) Clinical variables (two-year survival, site of cancer, histology of cancer, number of primary cancers); and (iii) Registration variables (method of diagnosis, number of notification episodes, type of notifying institutions).

A summary of the distribution of each variable by localised and unknown degree of spread is included in Appendix A. Based on initial analyses variables were re-categorised to try to

achieve the best consistency across time period and notification method. All continuous variables were categorised when non-linear associations with localised degree of spread were found to maximise consistency of association. An overview of the categorisation, association and consistency of each association for each variable is presented in table 1.

Based on univariate analyses, all variables except sex and number of primary cancers were significantly related to localised degree of spread. However, there appeared to be seven variables that maintained consistent association over time: age group; socio-economic status of area of residence (highest quintile vs lowest four quintiles); area of residence at diagnosis (metro vs non-metro); survival status (alive>2yr vs died <2yr); site of cancer (Lung& Bronchus NOS vs other sites); method of diagnosis; and number of notification episodes. Only three of these variables remained consistent across notification method (area of residence, site of cancer and number of notification episodes).

Building a multivariate logistic model

The predictive value of variables and consistency of association at a univariate level provides some indication as to which variables may be reliably included in a predictive multivariate model. However, variables can behave in a different manner when included in a multivariate model due to their associations with other covariates. Therefore, a similar process was used to build and assess a logistic regression model as for assessing univariate predictors. The aim was to build a model that would:

- Provide a reasonable level of prediction;
- Be a good fit of the data;
- Behave in a similar predictive manner independently of time period ;
- Behave in a similar predictive manner independently of notification method (electronic versus manual)

A “reasonable level of prediction” was assessed by examining the area under the ROC curve which indicates the combined sensitivity and specificity of the model. The aim was to achieve at least 70% prediction accuracy based on this measure.

A “good fit of the data” was assessed by examining the Adjusted R-square values and by using the Hosmer and Lemeshow Goodness of fit test.

Table 1: Association and consistency of association of variables with localised degree of spread

Predictor	Description	Categorisation	χ^2 ($p(\chi^2)>0$)	Consistent across time periods within manual notifications	Consistent across notification method within time period 3 (1999- 2004)
Demographic variables					
Age Group	Age at diagnosis, categorised into 3 age groups	<65 65-74 75+	215.1 ($p<0.01$)	yes	No
Sex	Sex	M F	0.28 ($p=0.60$)	yes	yes
Socio-economic status of residence at diagnosis	Based on SEIFA Index of Relative Disadvantage for postcode of residence at the time of diagnosis. Indexes were available for the years 1986, 1991, 1996, 2001, and 2006 and the closest index to the year of diagnosis was used. Data were grouped into quintiles.	Highest quintile Lowest four quintiles	64.5 ($p<0.01$)	yes	yes
Area of residence at diagnosis	Based on postcode of residence at time of diagnosis, cases are coded to one of the 8 Area Health Services (AHS) of NSW (2005 definition). The AHS were then re-grouped as Metropolitan (Hunter & New England, North Coast, Greater Southern, Greater Western) and Non-metropolitan (Sydney South West, South Eastern Sydney & Illawarra, Sydney West, Northern Sydney & Central Coast)	Metropolitan Non-metropolitan	237.3 ($p<0.01$)	yes	yes
ATSI status	Cases are coded as ATSI, Non-ATSI and unknown. ATSI and non-ATSI were grouped due to low case numbers	Known Not Known	48.8 ($p<0.01$)	no	no
Clinical variables					
Two year survival	All cause survival from time of diagnosis	Died <2 yrs Alive >2yrs	538.9 ($p<0.01$)	yes	no
Site of cancer	Based on ICD10 categorisation at a four digit level. Grouped as Lung & Bronchus NOS (C349) and all other sites within the Tumour Group "Lung" (C339, C340-C348).	Lung & Bronchus NOS All other sites	1145.7 ($p<0.01$)	yes	yes
Histology group	Based on ICD-03 categorisation. Small Cell cancers and all other non-specific codes grouped due to low case numbers.	Squamous Cell Carcinoma (SCC) Adenocarcinoma Small Cell and Other	110.8 ($p<0.01$)	no	No
Number of primary cancers	Number of primary cancers recorded at diagnosis	One Two or more	0.17 ($p=0.68$)	yes	yes
Registration variables					
Method of diagnosis	The method used to code clinical aspects of the cancer diagnosis. Other includes clinical, cytology, post-mortem reports but no histopathology. Cases based on death certificate only were excluded from the analyses.	Histopathology sighted at CCR Histopathology Other	1828.3 ($p<0.01$)	yes	no
Number of notification episodes	Count of notification episodes received for that case	One-two Three or more	657.8 ($p<0.01$)	yes	yes
Notifying facility type	Based type of notifying institution. Multiple episodes can be received from public and private hospitals, pathology labs, nursing homes, outpatient clinics. Grouped as cases including a private hospital notification and those not including a private hospital notification.	Includes private Does not include private	577.3 ($p<0.01$)	no	no

By “behaving in a similar predictive manner” it is meant that variables included in the model must have the same relationship with the outcome variable regardless of time period or notification method (broadly speaking this means they are always positive predictors or always negative predictors). Ideally all variables should also have roughly the same magnitude of prediction in all situations (indicated by similar sized regression coefficients). However, a reasonable prediction consistency was considered to be satisfied if the covariates had the same direction of relationship, with potentially differing strengths of relationship.

A full main effects logistic model (labelled Model 1) was initially assessed across 5 separate datasets: manually notified cases (M) for periods 1, 2, and 3; and for Electronic only (E) and Electronic plus manual (EM) cases within period 3. The results of these regression analyses are presented in tables B1 and B2 in Appendix B.

Five variables were significant predictors after controlling for other factors and behaved reasonably consistently across time periods: age; area of residence at diagnosis; histology; site of cancer; survival; and method of diagnosis. Four of these five were identified as consistent at a univariate level with histology now appearing as reasonably consistent across time after controlling for other factors. The number of notification episodes no longer appeared consistent after controlling for other factors.

Area of residence at diagnosis, and site of cancer remained consistent across notification methods.

As notification method appeared to interact with most variables in predicting localised degree of spread, the decision was made to just focus on data available in period 3, in which all three notification method groups were available with no missing data. This allows models to take into account differences across notification methods. However, given that the missing data exist in a distinct time period, only variables shown to behave consistently across time periods were considered for analysis. Two further models were assessed:

- Model 2: a reduced main and interaction effects model including only significant and consistent variables as main effects and significant interaction terms. This model was applied separately to E and EM cases in period 3.
- Model 3: the same as model 2, but including notification method as a main and interaction effect. This model was applied to a combined dataset of all E and EM cases in period 3.

Proc Logistic within SAS version 9.2 was used for all analyses.

Results

Table 2 and Figure 5 provide a summary of the results from Model 2 applied separately to E and EM notified cases within period 3. These models were shown to have reasonable predictive power, although they were borderline for meeting the predictive benchmark established *a priori*. The model for EM notified cases was slightly more predictive with 71.5% prediction accuracy based on the area under the ROC curve. This compared to 67.0% prediction accuracy for E notified cases. The model also provided a better fit of the data for EM cases compared to E only. Both provided an adequate fit of the data based on the Hosmer and Lemeshow Goodness of Fit test.

Table 3 provides a summary of results from Model 3 for the E and EM combined dataset for period 3. Only one interaction effect with notification method remained significant (method of diagnosis by notification method) so others were excluded. This model appeared to have good predictive power at 76% based on the area under the ROC curve and was an adequate fit of the data.

Direct comparison of the model diagnostics between Models 2 and 3 is difficult given that model 3 is based on a larger combined dataset.

Both Models 2 and 3 appear to provide reasonable prediction for localised versus unknown degree of spread and fit the data adequately. While Model 3 is more parsimonious, both were applied within a multiple imputation procedure to test the viability of this process. Two major assumptions are made in applying these models to impute the missing data:

- 1) The variables included within the model have the same predictive effect within E and EM cases in 1993-1998 as they do within 1999-2004
- 2) The underlying proportions of localised and unknown cases in 1993-1999 are similar to those in 1999-2004.

There is no direct way to test these two assumptions. However, to support the validity of assumption 1, only variables consistent across periods based on the manual notification group were included. Assumption 2 appears reasonable based on the finding that the proportions were similar for the periods before and after the 1993 to 1998 period.

Table 2: Logistic Regression results for predicting localised lung cancer: Model 2 applied to a) E notified cases and b) EM notified cases in period 3

Parameter	Reference category	Model 2(a)		Model2(b)	
		Estimate	Pr > ChiSq	Estimate	Pr > ChiSq
Intercept		-0.95	<0.01	0.51	<0.01
Histology	Adenocarcinoma	-0.41	0.02	0.15	0.16
	Sm. Cell and other	-0.46	<0.01	-0.27	0.01
Method	Histo sighted at CCR	-1.36	0.08	0.78	<0.01
	Histopathology	0.41	0.00	0.63	<0.01
Site	Bronchus & Lobes NOS	0.83	<0.01	-0.72	<0.01
AHS	Non-Metro	-0.26	0.02	-0.62	<0.01
Survival	alive > 2yr	-0.85	<0.01	-0.08	0.80
Survival*Method	alive > 2yr*Histo-CCR	2.77	0.01	1.24	0.00
	alive > 2yr* Histopathology	1.11	<0.01	0.45	0.19
Adj r squared		0.11		0.160	
% Concordance		63.9		69.4	
% Discordance		30.3		26.3	
Goodness of fit		10.68	0.15	6.40	0.70

Figure 5: ROC curve for Model 2 applied to a) E notified cases and b) EM notified cases in period 3

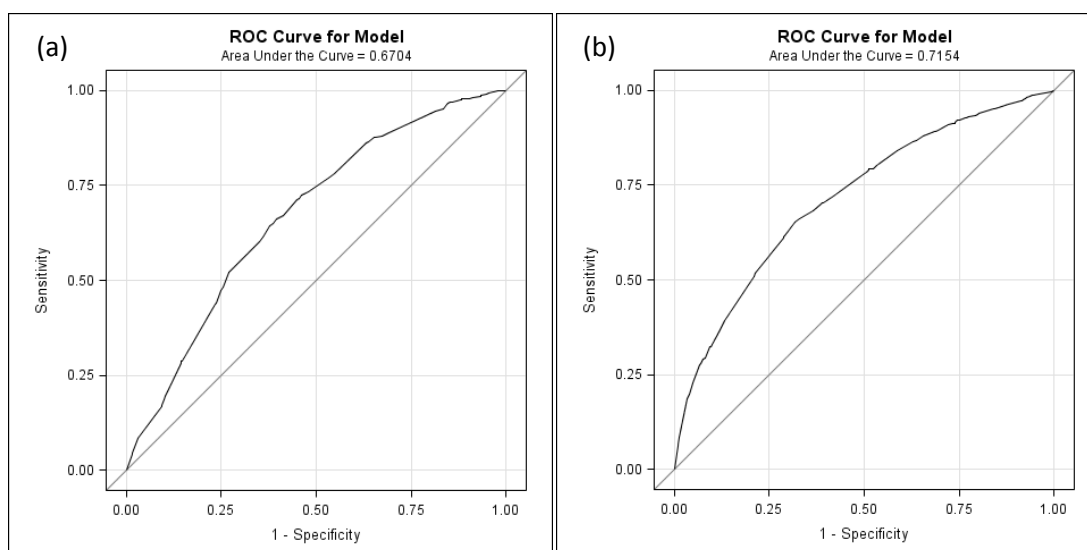
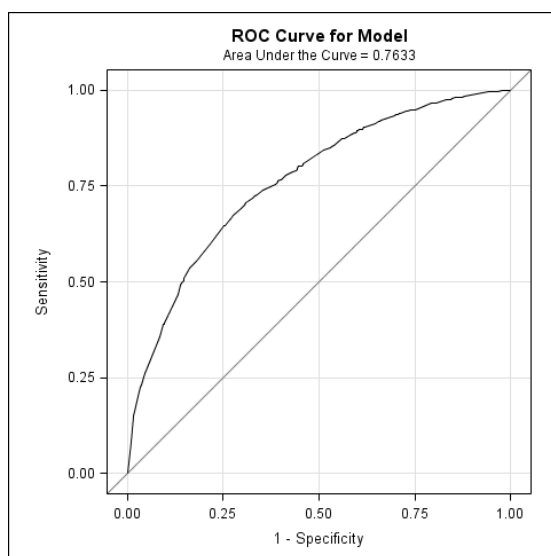


Table 3: Logistic Regression results for predicting localised lung cancer: Model 3 applied to E and EM notified cases in Period 3

				Model 3
				E and EM
Parameter		Reference category	Estimate	Pr > ChiSq
Intercept			0.58	<0.01
Histology	Adenocarcinoma	SCC	-0.01	0.90
	Sm. Cell and other	SCC	-0.34	<0.01
Method	Histo sighted at CCR	Other	0.70	<0.01
	Histopathology	Other	0.58	<0.01
Site	Bronchus & Lobes NOS	Other	-0.80	<0.01
AHS	Non-Metro	Metro	-0.48	<0.01
Notification method	E	EM	-0.70	<0.01
Survival	alive > 2yr	Died<2yr	-0.52	0.01
Survival*Method	alive > 2yr*Histo-CCR	Died<2yr*Other	1.72	<0.01
	alive > 2yr* histopathology	Died<2yr*Other	0.84	<0.01
Notification method*Method	E*Histo-CCR	EM*Other	-1.81	<0.01
	E* histopathology	EM*Other	-0.18	0.26
Adj r squared			0.27	
% Concordance			75.3	
% Discordance			22.6	
Goodness of fit			12.43	0.19

Figure 6: ROC curve for Model 3 applied to E and EM notified cases in Period 3



Section 3: Multiple Imputation to correct the degree-of-spread data artefact for lung cancer cases

Imputation Procedure

The multiple imputation procedure within SAS 9.2 was used to produce new datasets based on the two models being tested: Model 2 as specified in table 2 and Model 3 as specified in table 3. Five imputations were undertaken on missing data by specifying each model using the logistic option within the MI procedure. For both models, the missing data were defined by creating a new variable for degree of spread which included original degree-of-spread values but was deemed 'missing' if:

- the value of stage was 'unknown'; and
- the notification method was electronic or electronic and manual; and
- the year of diagnosis was between 1993- 1998.

A further restriction was placed on the MI procedure: the 'prototype' cases to be used for predicting the missing data were limited to only those cases diagnosed in the period 1999-2004 by electronic only (E) or electronic plus manual (EM) means. To produce the imputed data using the MI procedure in SAS, only the prototype cases and the cases with missing degree-of-spread data were included.

The MI process produces five datasets for each model representing the five iterations of the imputation process. An analysis dataset is created for each model which combines the five datasets and Proc MIAnalyze was used to combine estimates from the five imputed datasets.

Results - Imputation

Based on original coding the proportion of localised cases of lung cancer (within all localised and unknown) diagnosed in 1993-1998 was 32.0%. Following multiple imputation the proportion was 50.6% with a 95% Confidence Interval of 49.3 - 52.0%.

Table 4 shows the proportion of unknown cases re-coded to localised following imputation. There was an increase in recoding from 1993-1997 which reflects the rapid increase in frequency of electronic notifications. The proportion of cases recoded by year was similar for both models.

Table 4: Percentage of unknown cases re-coded to localised following imputation, by year of diagnosis

		1993	1994	1995	1996	1997	1998
Model 2	%	21.9	26.2	28.0	29.8	30.3	26.7
Model 3	%	22.5	26.9	27.4	29.6	31.2	26.9

Figure 7 shows the percentages localised and unknown prior to any imputation and following imputation based on Model 2. The model appears to correct the artefact. Figure 8 shows the same following imputation based on Model 3. This model also appears to correct the artefact, following a very similar trend as per Model 2. Figure 9 compares the two models, and shows that the results are almost identical.

Given the similarities between the two models, Model 3 was the more parsimonious and was the preferred model for further investigations.

Figure 7: Model 2 – percentage localised and unknown pre and post imputation

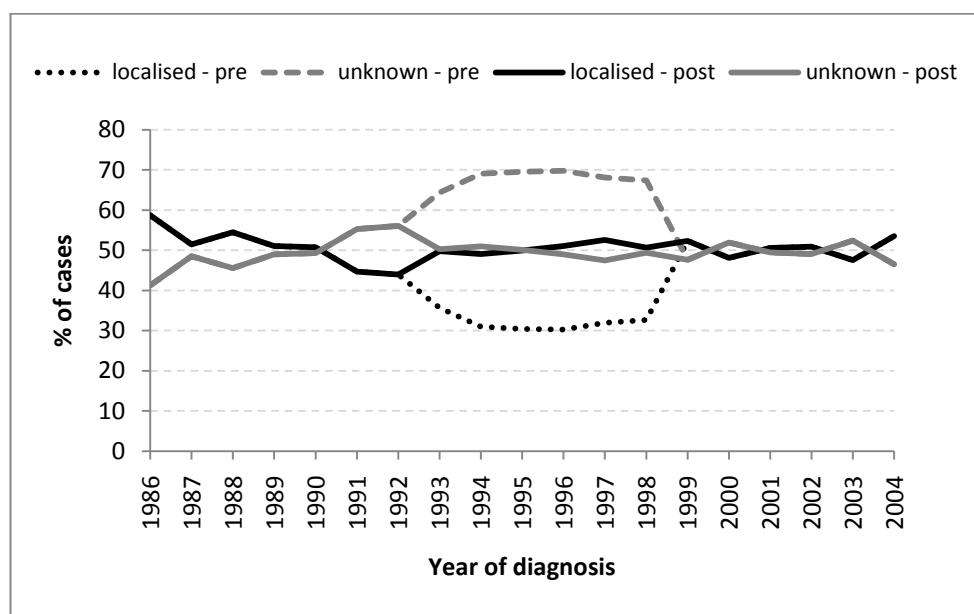


Figure 8: Model 3 – percentage localised and unknown, pre and post imputation

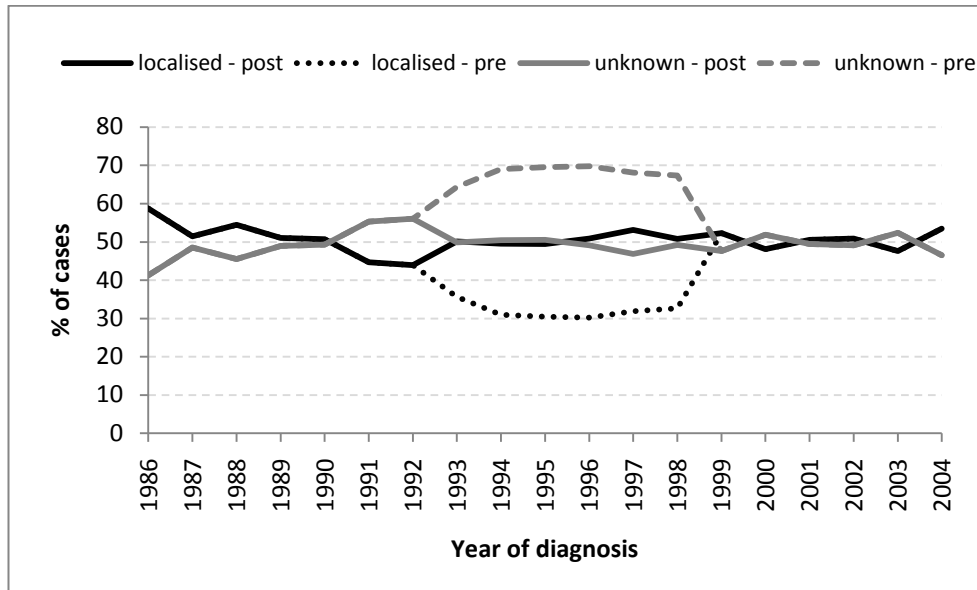
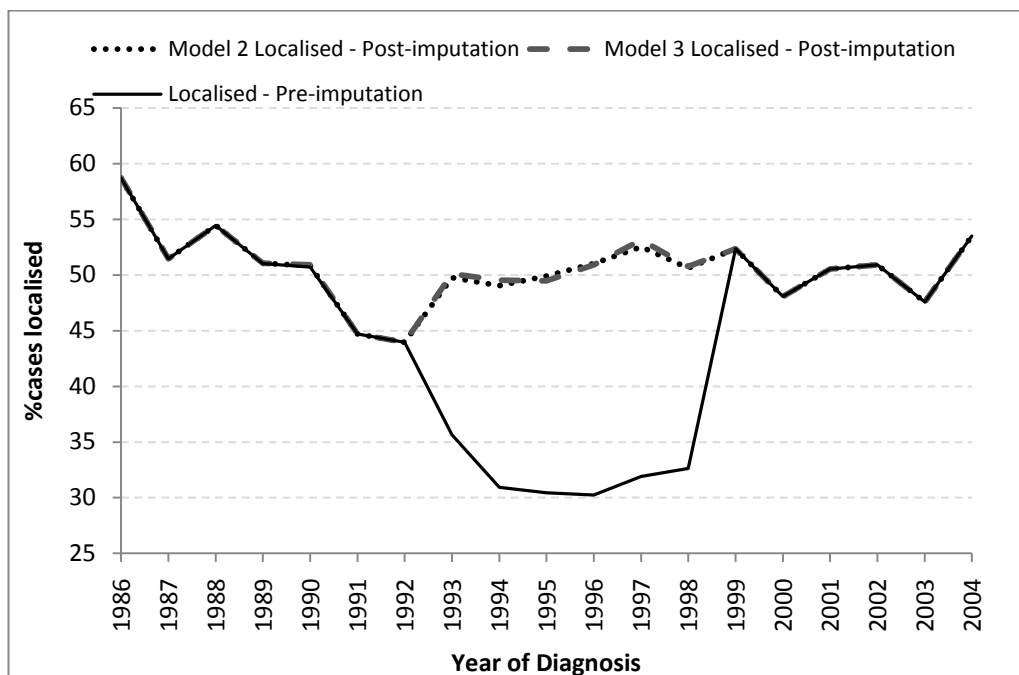


Figure 9: Percentage localised pre- imputation and post- imputation, Models 2 and 3



Validation of the MI model

The results produced using MI are plausible in terms of correcting the data artefact. However, due to the missing data pattern, some major assumptions were required to implement this procedure. The missing data for degree of spread were for *all* E and EM cases within a distinct time period. Additionally, method of notification appeared highly related to degree of spread. This meant that manually notified cases within the time period 1993-1998 were not deemed appropriate cases upon which to base a model predicting missing data in E and EM cases.

Instead cases within a distinct time period for just E and EM cases were used. In doing so, we are assuming that the relationships between predictors and degree of spread within the E and EM groups are the same within the 1993-1998 period as they are within the 1999-2004 period. We are also assuming that the ratio of localised to unknown cases is similar in the two periods.

Two additional sets of analyses were undertaken to assess the validity of the imputation model. Firstly, the effect of imputation by covariates was examined to identify any potential biases introduced within sub-groups. Secondly, the same MI process using 1999-2004 cases was applied to predict all localised and unknown E and EM data in the period 2005-2006. This enabled direct comparison between actual values and predicted values. While it does not directly assess the assumptions made for the period 1993-1998 it provides an alternate mechanism for testing the validity of applying the model in a different time period.

Results - validation

Imputed data patterns by covariate

The effect of imputation by covariates is presented for Model 3 in figures 10(a)-(f). The effect of imputation appears consistent across all levels of covariates with the exception of Method of diagnosis = 'Other'. For this grouping there appears to be a possible 'over-compensation' with an increase in proportion localised during the artefact period. However, closer analysis of this grouping by method of notification (figure 11) shows that there was an increase in proportion localised for manually notified cases also, which suggests that other factors may have influenced this trend.

Figure 10a: Model 3 – pre and post imputation by NOTIFICATION METHOD

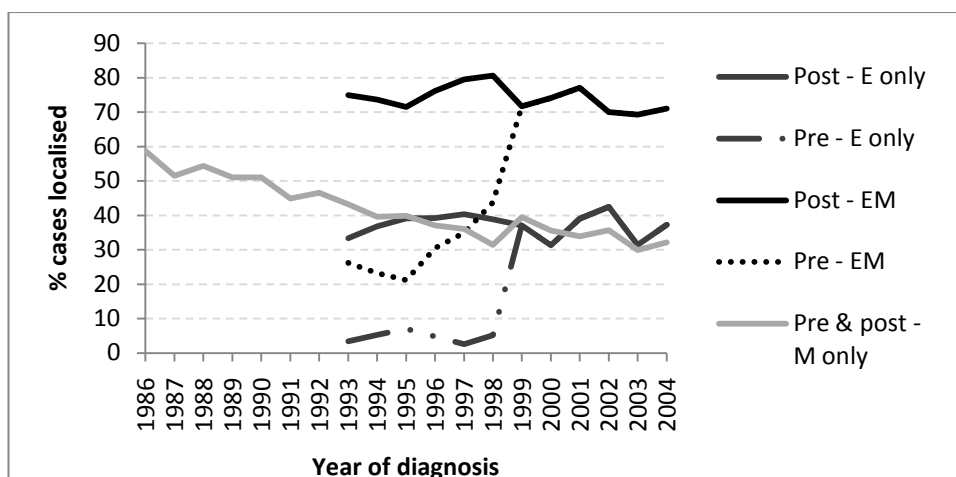


Figure 10b: Model 3 – pre and post imputation by METHOD of DIAGNOSIS

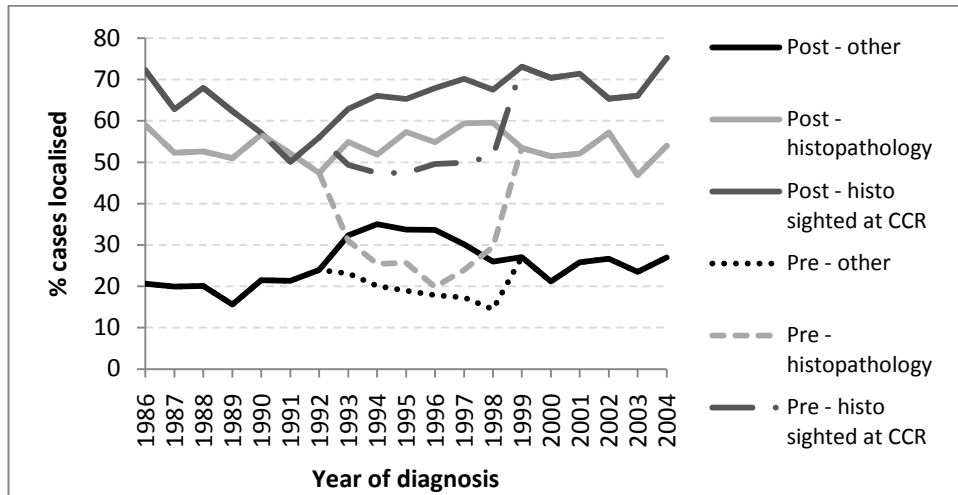


Figure 10c: Model 3 – pre and post imputation by SURVIVAL

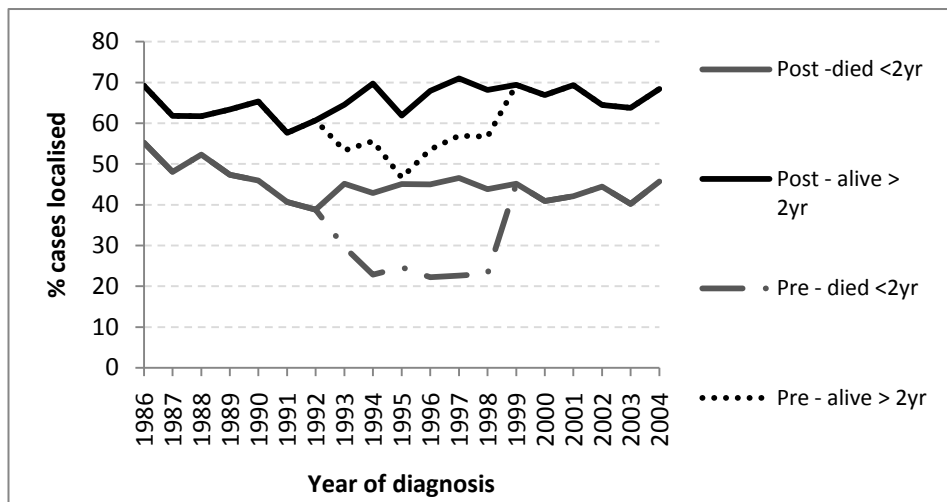


Figure 10d: Model 3 – pre and post imputation by HISTOLOGY

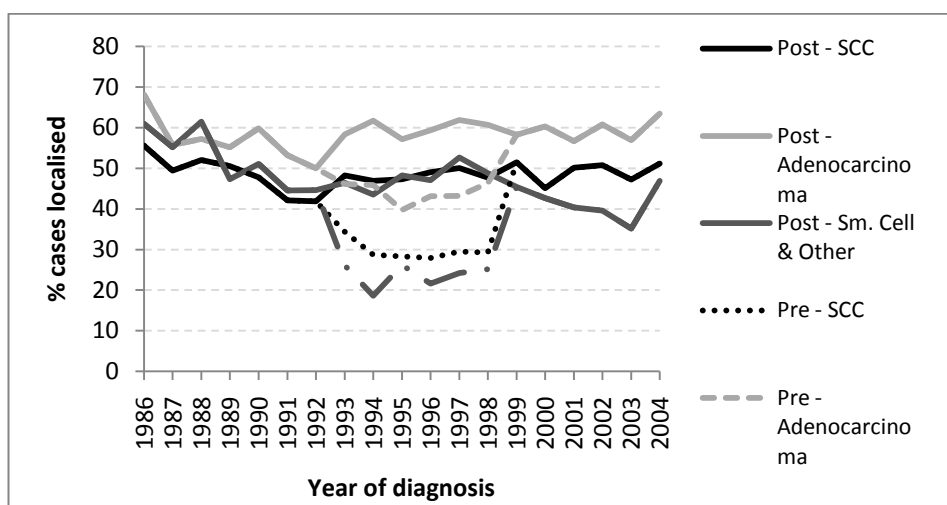


Figure 10e: Model 3 – pre and post imputation by SITE of LUNG CANCER

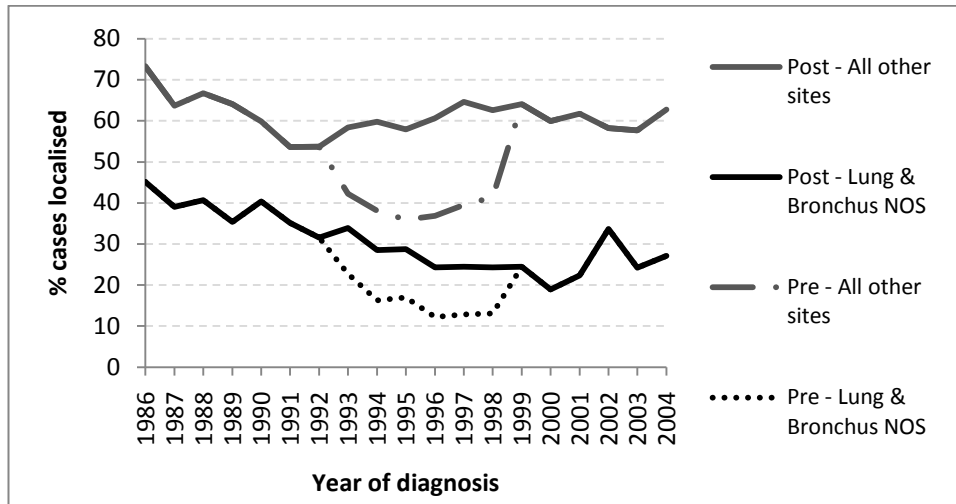


Figure 10f: Model 3 – pre and post imputation by AREA

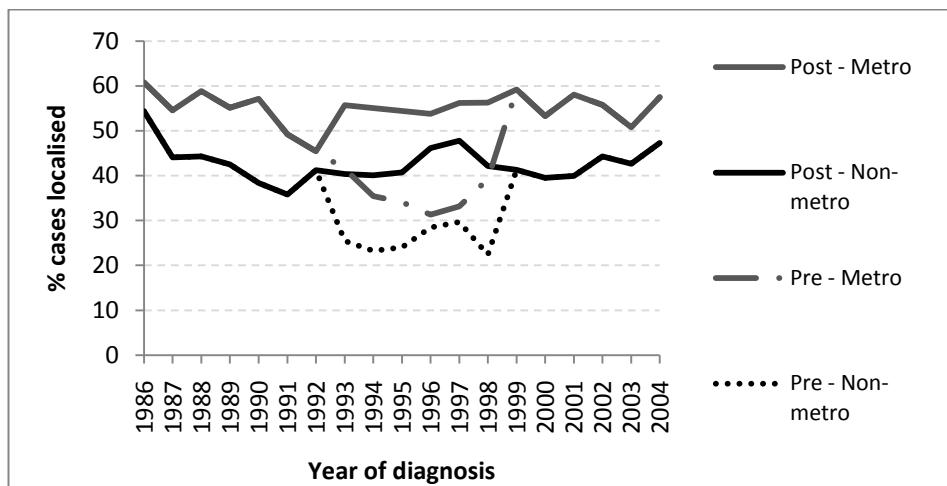
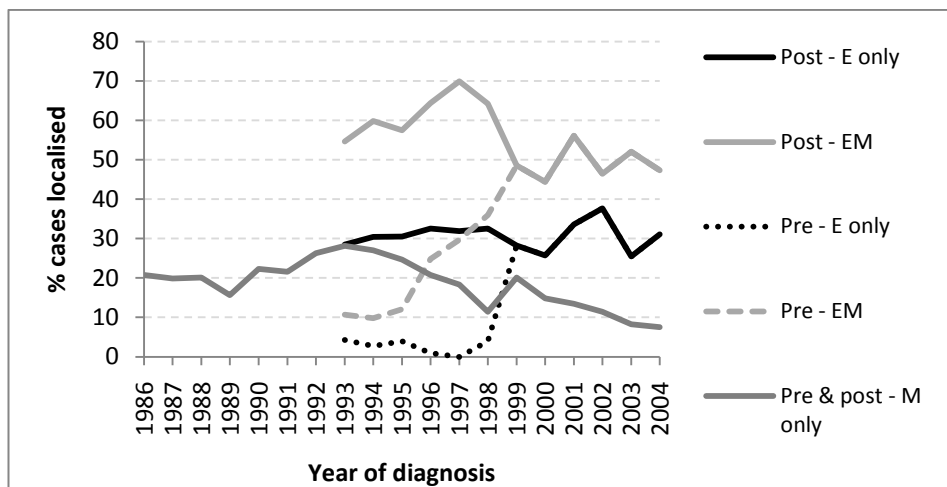


Figure 11: Model 3 – pre and post imputation by NOTIFICATION METHOD for METHOD of DIAGNOSIS = Other



Predicting cases in a distinct time period

Two scenarios were tested using cases diagnosed within 2005-2006. Firstly, a sample of 1,000 test cases was randomly selected from all E and EM notified localised and unknown cases within this period. These test cases reflected the underlying proportion of 53.8% localised versus 46.2% unknown degree of spread for these years. This was a different proportional split from the 'prototype' cases within the period 1999-2004 in which 60.7% were localised. Secondly, a sample of test cases was purposefully selected to ensure the proportion localised matched that of the prototype cases (with 60.7% localised).

The test cases were designated as having missing data for degree of spread. Data was then multiply imputed for the degree-of-spread variable for test cases and imputed values compared with original 'known' values.

Table 5 indicates the effectiveness of the MI model in correctly predicting cases of unknown and localised cases (E and EM notified) for the period 2005-2006 based on scenario 1. Prediction was reasonable for localised cases, with the model correctly predicting 69.0% with a 95% confidence interval of 63.7-74.4% based on the multiply imputed datasets. In terms of predicting localised cases, this shows the model as having high sensitivity. However, the model was poor for predicting unknown degree-of-spread cases, only correctly predicting 48.3% (95% CI: 42.8-52.8%) showing poor specificity.

Table 5: prediction of E and EM notified, localised and unknown cancers for 2005-2006

		Original Coding		
		Localised	Not Localised (Unknown)	All cases
Test Outcome	Localised	371	239	610
	Not Localised (Unknown)	167	223	390
	All cases	538	462	1000
Sensitivity (prediction of localised)		69.0%		
Specificity (prediction of unknown)		48.3%		

Table 6 presents the results from this analysis based on scenario 2. For scenario 2, the data is missing at random, as the probability of a case being localised within the sample is independent of whether the missing data are included or excluded. The predictive power for localised cases is 69.5% (95% CI: (63.1-76.0)), which is similar to that found within scenario 1. Again, the specificity is poor at 48.2% (95% CI: 40.0-59.4).

Table 6: Prediction accuracy for E and EM notified, localised and unknown cancers – 2005-2006, sample meeting the MAR assumption

		Original Coding		
		Localised	Not Localised (Unknown)	All cases
Test Outcome	Localised	422	204	626
	Not Localised (Unknown)	185	190	375
	All cases	607	394	1000
Sensitivity (prediction of localised)		69.5%		
Specificity (prediction of unknown)		48.2%		

Overall, the model appeared to predict localised cases with reasonable sensitivity in a distinct time period. However specificity of the model was poor. These data suggest that the model performs moderately in a distinct time period but does not perform as well as it does in the 1999-2004 period on which it is based. The results of this validation showed that prediction accuracy was not affected to a large extent by differing proportions of localised versus unknown degree of spread in the test period.

The results of this validation suggest that caution should be exercised in applying this procedure in the manner described and the assumptions inherent in the process should be clearly articulated.

Section 4: Impact of correcting the data artefact on survival estimates

Due to the relationship between survival and degree of spread, it is likely the data artefact will have impacted upon this type of analysis. To assess whether the data artefact could impact survival estimates, trends in survival by time of diagnosis were examined using the original pre-imputation data.

For localised cases, Kaplan-Meier survival curves were computed by three year period based on original degree-of-spread coding (Figure 12a). The two time-of-diagnosis periods within the artefact period – 1993-1995 and 1996-1998 – had the third lowest and lowest mortality rates, with median survival of 20 and 27 months respectively. This is out of step with the overall pattern of survival across time periods, where there is evidence of a trend towards increasing survival over time with median time to death of 12, 13, 18 and 21 months for the periods 1986-89, 1990-92, 1999-01, and 2002-04 respectively. For regionalised cancers, which were unaffected by the artefact, a similar trend in increasing survival is evident across all six time periods (Figure 13).

Post-imputation, the survival trend across time for localised cancers appears more in accordance with that seen for regionalised cancers, with survival increasing uniformly with each time period (Figure 12b). Comparing the survival curve for localised cases pre and post imputation for the artefact period (1993-1998), we can see that survival is significantly lower following imputation (Figure 14a). Based on originally coded data, median time to death was 23 months (95% CI 22-24), which reduced to 15 months (95% CI 15-16) based on imputed data.

Survival from lung cancers diagnosed with 'unknown' degree of spread was not similarly affected by the data artefact, with very little difference in survival evident across time periods and imputation had little impact on estimates for the artefact period (Figure 14b). Median survival from unknown lung cancer diagnosed within the period 1993-1998 was 9 months based on original data and 8 months based on imputed data, which was not significantly different ($p=0.23$).

The reason for the difference in impact of imputation for localised cancers compared to unknown cancers can be explained by the differences in survival by degree of spread and notification method. Based on data from 1999-2004 for localised cancers diagnosed within the M group, survival appears substantially better than for both the E and EM groups (Figure 15a). However, for cases with unknown degree of spread, survival differences by notification method were substantially less (Figure 15b). This means that when cases from the E and EM groups are re-categorised from unknown to 'localised' they tended to be cases with worse

survival than those in the M group and therefore reduced the survival rate of the group as a whole.

This may be linked to the fact that pathology reports (which are manual notifications) are generally received for lung cancer when surgery has been undertaken, and that surgery for localised lung cancer is an indicator of better outcomes (7).

Figure 12a : Kaplan-Meier survival curve for lung cancer cases with localised degree of spread, by period of diagnosis, PRE-IMPUTATION

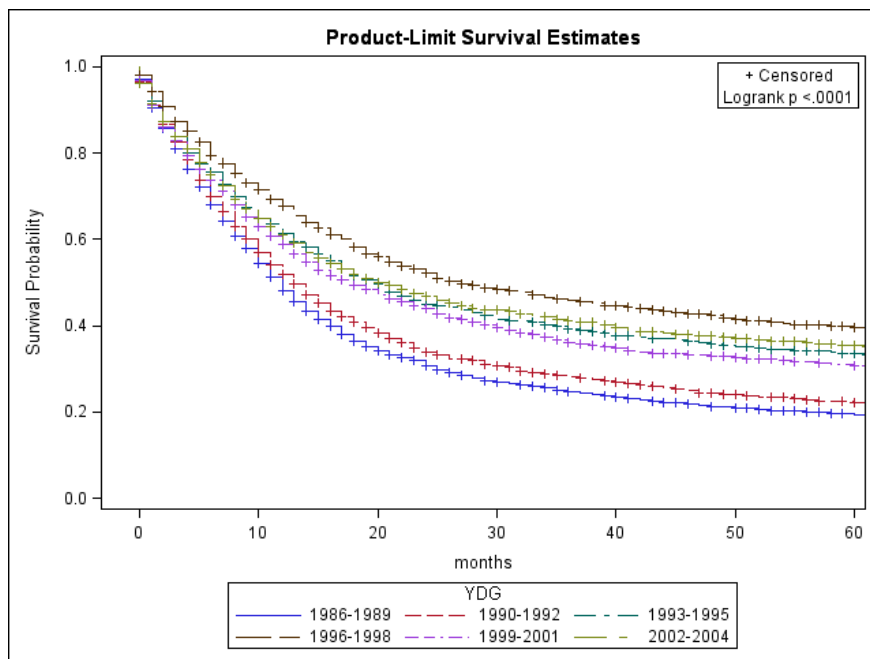


Figure 12b: Kaplan-Meier survival curve for lung cancer cases with localised degree of spread, by period, POST-IMPUTATION

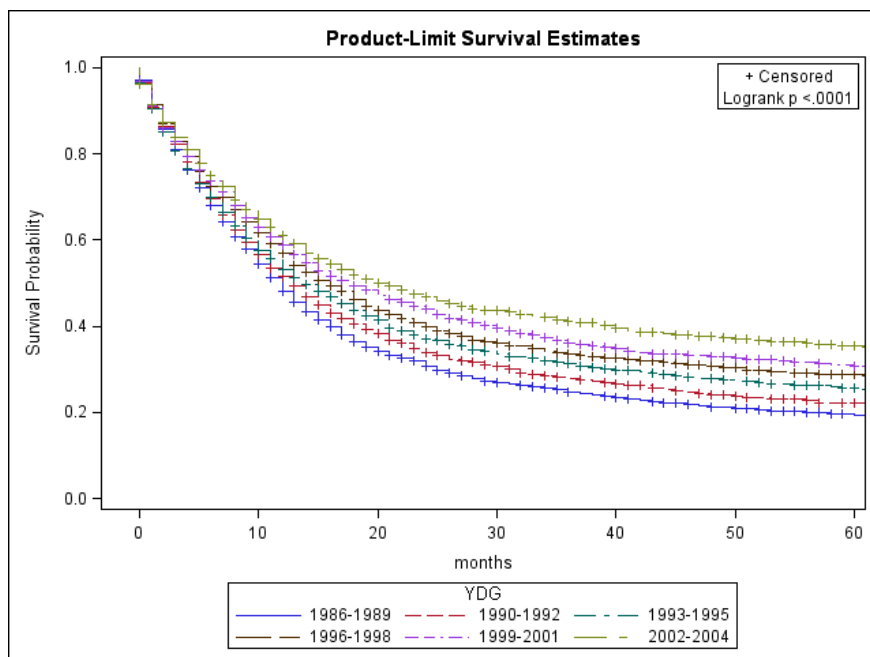


Figure 13: Kaplan-Meier survival curve for lung cancer cases with regionalised degree of spread, by period of diagnosis

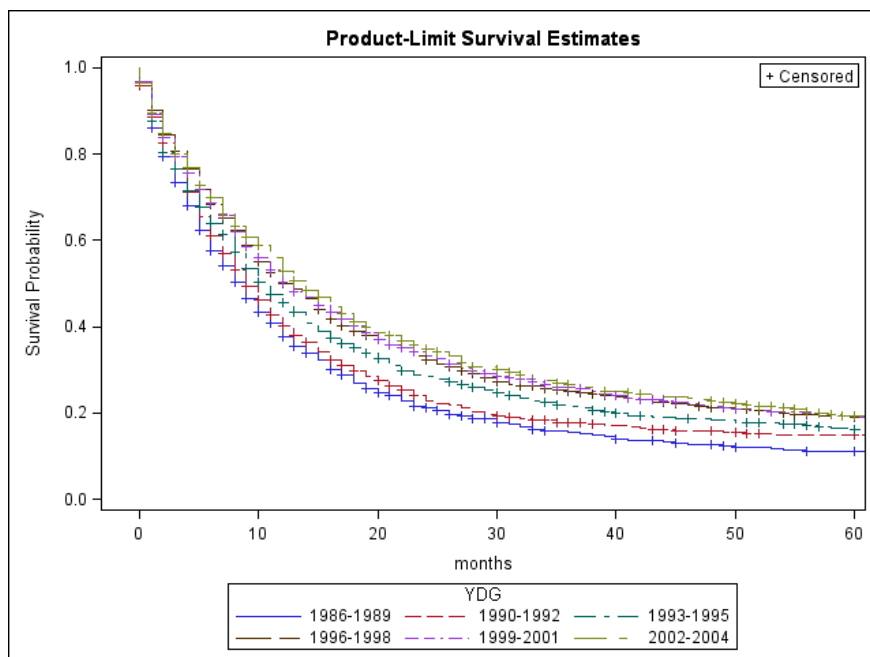


Figure 14a : Kaplan-Meier survival curve for lung cancer cases diagnosed 1993-1998 with localised degree of spread, Pre- and Post- imputation

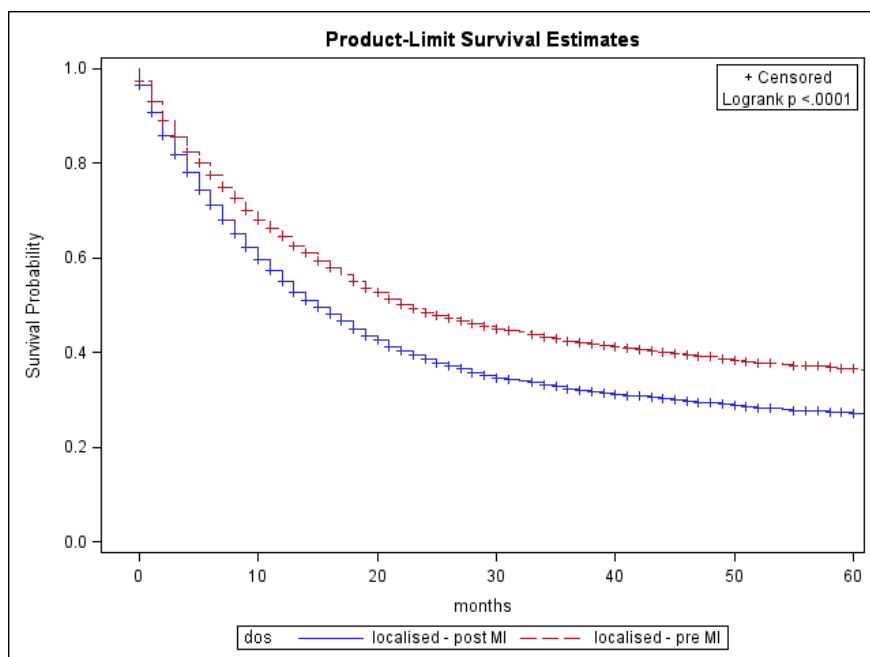


Figure 14b : Kaplan-Meier survival curve for lung cancer cases diagnosed 1993-1998 with unknown degree of spread, Pre- and Post- imputation

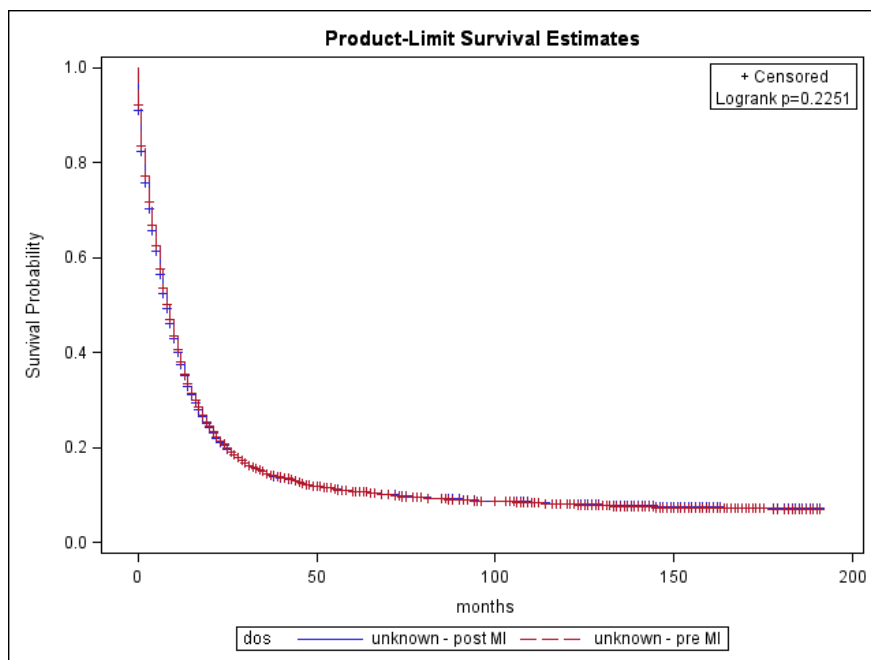


Figure 15a : Kaplan-Meier survival curve for lung cancer cases diagnosed 1999-2004 with localised degree of spread, by notification method

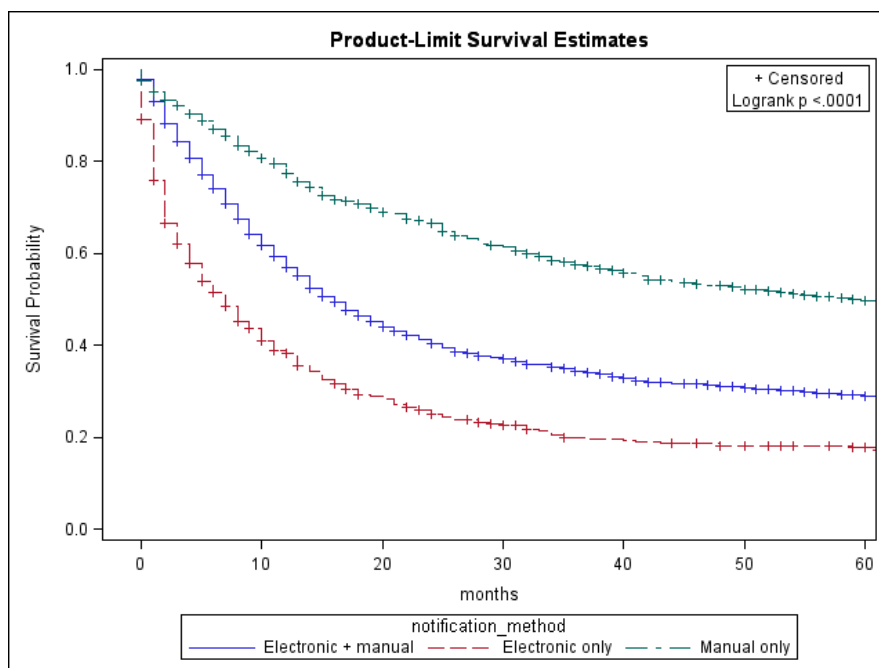
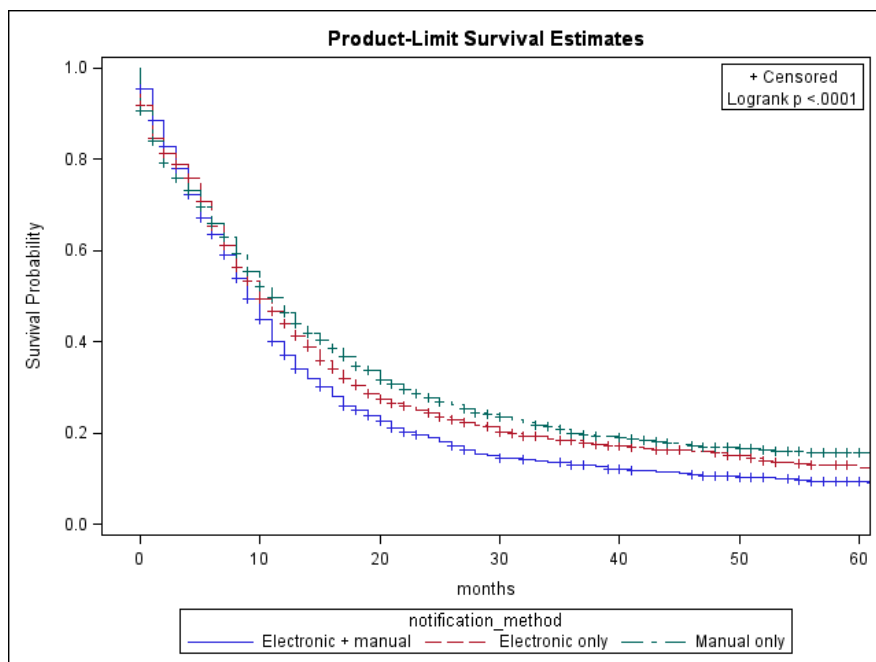


Figure 15b : Kaplan-Meier survival curve for lung cancer cases diagnosed 1999-2004 with unknown degree of spread, by notification method



Section 5: Discussion and Conclusions

The MI procedure using the logistic option based on model 3 appears to provide a reasonable approach to correcting the data artefact for degree of spread for lung cancer. Plausible results were produced that are consistent across levels of other covariates in the database. This model requires that cases are categorised into groups based on method of notification, but aside from this relies only on readily available variables within the CCR reporting database.

The data artefact that currently exists for the period 1993-1998 appears to significantly impact on survival from localised lung cancer for this period. Correction of the data artefact based on the imputation process used here produces significantly lower survival estimates for localised lung cancer cases within the 1993-1998 period; however, survival trends by time of diagnosis appear more consistent. The imputation process did not appear to impact on survival estimates for unknown cases.

In order to implement this procedure, cases recorded originally as having unknown degree of spread were designated as “missing” for all Electronic (E) and Electronic plus manual (EM) notifications within a distinct time period (1993-1998) based on year of diagnosis. This meant that the predictive model had to be based either on cases notified by a different means (manual only), or on cases notified within a different time period (1999-2004). Given the observed differences in ratio of localised-to-unknown cases for different methods of notification, the chosen approach was to use cases from a different time period for E and EM notifications only. Two major assumptions are therefore made: the overall ratio of localised to unknown cases is similar across time periods; and the relationship between predictors and localised degree of spread within the E and EM groups are the same within the 1993-1998 period as they are within the 1999-2004 period.

Validation of the chosen imputation process using cases from 1999-2004, but applied to cases diagnosed in a independent time period (2005-2006) with known degree-of-spread coding, suggested that the model still had an adequate ability to predict localised cases. However, the results suggested that some sensitivity and specificity was lost since model parameter estimates developed from the 1999-2004 period were applied to 2005-06.

The MI procedure appears a useful approach to correcting the known data artefact for degree of spread in lung cancer. However, given the major assumptions made and level of prediction achieved, it is recommended that caution is applied in using the imputed data. While values appear plausible when examined holistically and across broad levels of covariates, it is not considered appropriate to use the data at a case level. For analyses of small sub-groups

further investigation of the validity of this model would be required. In all instances, appropriate methods should be used to account for the increased error that is introduced via imputation.

This study describes the application of this procedure to one cancer type – lung cancer. The data artefact in the NSW Central Cancer Registry is known to exist to varying degrees for all solid tumours, excepting breast and melanoma. The model used here includes some variables that are specific to lung cancer, including ‘site of cancer’ and ‘histology’ of cancer. For both these variables there appeared to be a relationship between categories that were non-specific such as ‘not otherwise specified’ and the increased likelihood of ‘unknown’ degree of spread. For this reason, it is likely that these variables will remain useful predictors for other cancer types, but categories would need to be specified on a cancer-by-cancer basis.

References

1. Barraclough, H., Morrell, S., Arcorace, M., McElroy, H., Baker, D.F. (2008) *Degree-of-spread artefact in the New South Wales Central Cancer Registry*. Australian and New Zealand Journal of Public Health, Vol. 32, pp. 414-416.
2. International Agency for Research on Cancer WHO, & International Association of Cancer Registries In D. Esteban, S. Whelan, A. Laudico, & D. Parkin (Eds.). (1995) *Manual for Cancer Registry Personnel, IARC Technical Report No 10*. Lyon
3. Schafer JL, Graham JW. (2002) *Missing data: Our view of the state of the art. Psychological methods*. Vol. 7, pp. 147-77
4. James R. Carpenter, Michael G. Kenward and Ian R. White. (2007) *Sensitivity analysis after multiple imputation under missing at random: a weighting approach*. Stat Methods Med Res, Vol. 16, pp. 259-275.
5. Rubin, D.B. (1996) *Multiple Imputation After 18+ Years*. Journal of the American Statistical Society, Vol. 91, pp. 473-489.
6. Tracey E, Kerr T, Dobrovic A, Currow D. (2010) *Cancer In NSW: Incidence and Mortality Report 2008*. Cancer Institute NSW. Sydney
7. Dominioni L, Imperatori A, Rovera F, et al. (2000) *Stage I nonsmall cell lung carcinoma: analysis of survival and implications for screening*. Cancer, Vol. 89, pp. 2334-2344.

Appendix A: Summary of potential predictors for degree of spread

Demographic variables

Age

Age **was significantly associated** with unknown versus localised degree of spread. The mean age of localised cases was 68.3 years which was lower than for cases with unknown degree of spread (70.8 years) ($t=-19.3$; $p<0.01$). Age did not appear to be linearly related to degree of spread with those in the 75+ age group showing much lower levels of localised cancer than those in younger age groups. The effect of age appeared reasonably consistent across time, but not across notification methods. Due to the non-linear association, age was treated as a categorical variable for further analyses.

Figure A1: Age distribution for localised and unknown cases

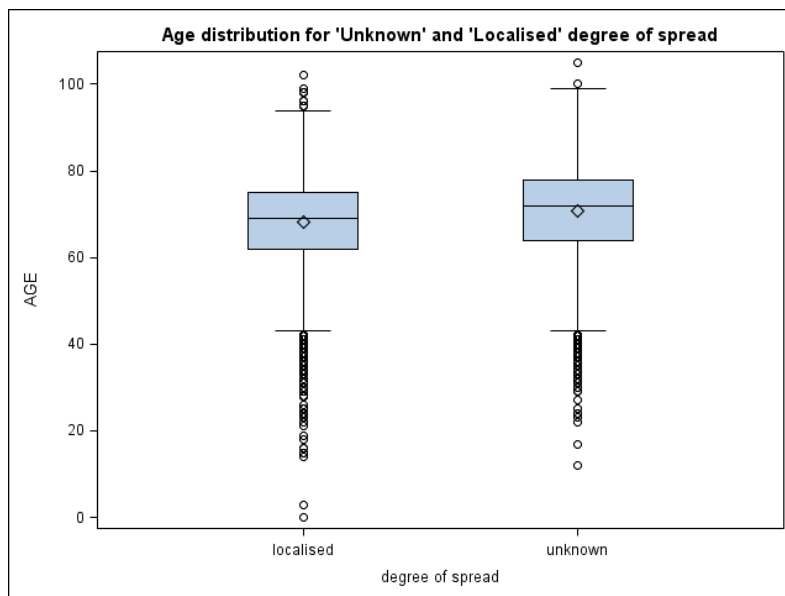
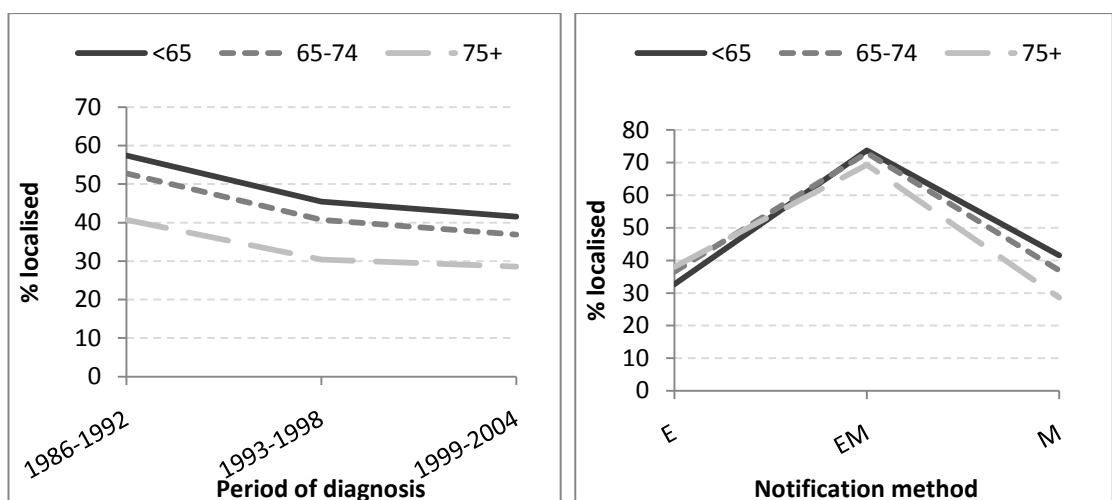


Figure A2: proportion localised by age group: (a) by period of diagnosis for M notified cases; (b) by notification method within period 3



Sex

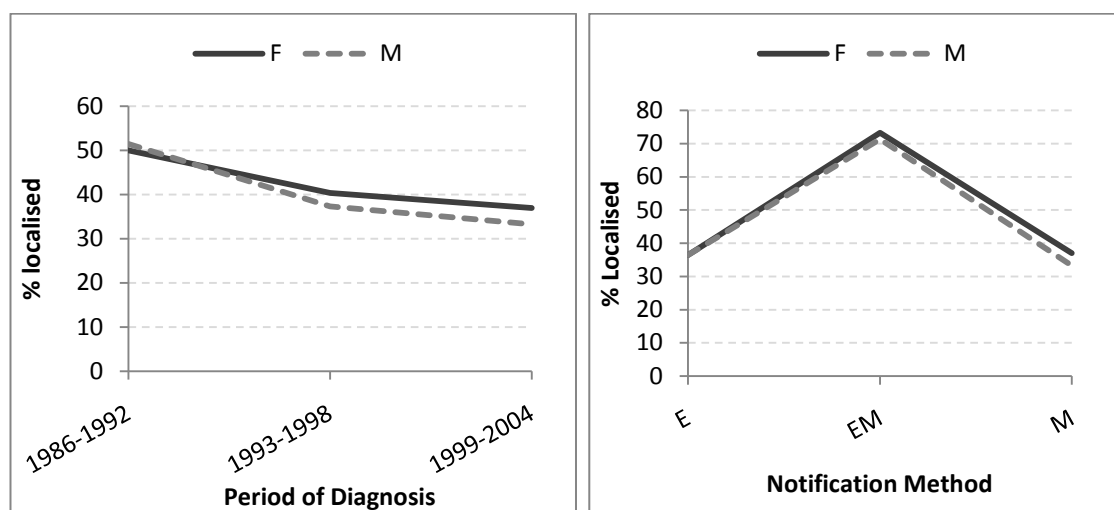
Sex of patient **was not significantly associated** with degree of spread of lung cancer at diagnosis ($\chi^2 = 0.28, p=0.60$).

Table A1: Distribution of Localised and Unknown cases, by sex

Sex		Localised	Unknown	Total (localised + unknown)	P value (χ^2)
Female	N	2567	2546	5113	0.60
	%	50.2	49.8		
Male	N	5981	5828	11809	
	%	50.6	49.4		

There was very little difference between males and females in all periods and across all notification methods.

Figure A3: proportion localised by sex: (a) by period of diagnosis for M notified cases; (b) by notification method within period 3



Socio-economic status

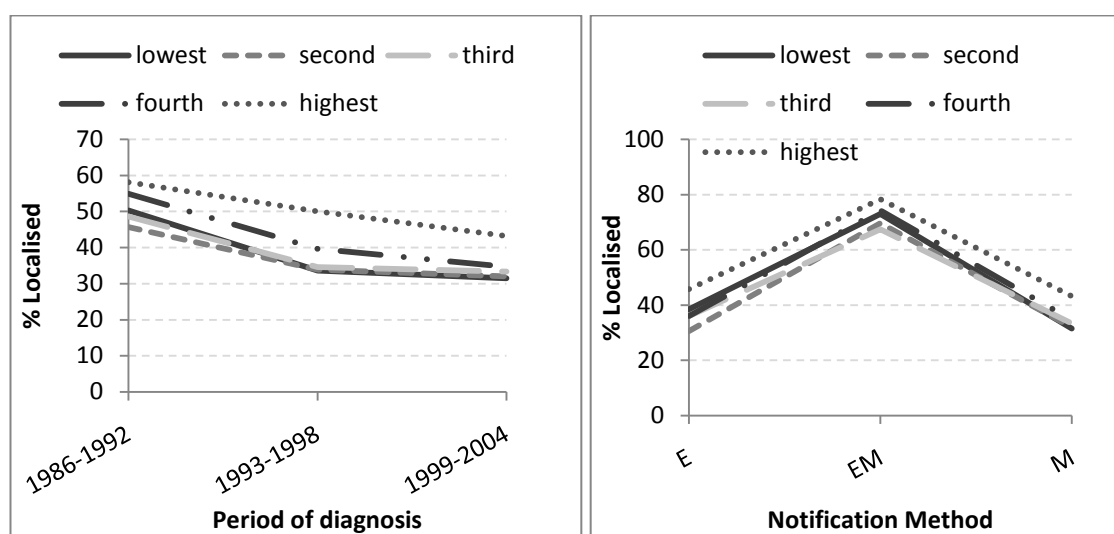
Socio-economic status (SES) was measured by SEIFA Index of Relative Disadvantage for postcode of residence at the time of diagnosis. Indexes were available for the years 1986, 1991, 1996, 2001, and 2006 and the closest index to the year of diagnosis was used. Data was grouped into SEIFA quintiles. SES showed a significant association with degree of spread ($\chi^2 = 98.69, p<0.01$). The proportion of cases with unknown degree of spread was highest in the second lowest SES quintile and tended to lower proportions in the higher SES quintiles. There were also a small number of cases for which SES was undetermined due to missing postcode information.

Table A2: Distribution of Localised and Unknown cases, by socio-economic status

IRSAAD quintile		Localised	Unknown	Total (localised + unknown)	P value (χ^2)
Lowest	N	1960	1962	3922	<0.01
	%	50.0	50.0		
Second	N	1781	2058	3839	
	%	46.4	53.6		
Third	N	1682	1825	3507	
	%	48.0	52.0		
Fourth	N	1579	1392	2971	
	%	53.1	46.9		
Highest	N	1507	1103	2610	
	%	57.7	42.3		
Unknown	N	39	34	73	
	%	53.4	46.6		

When considering manual notifications only, the association between SES and degree of spread was reasonably consistent across periods. The majority of difference in percentage localised appears to be occurring for the highest quintile compared to the lowest four. There was little difference between the lowest three quintiles in all periods and, within period 3, there was little difference between the lowest four quintiles across all notification methods. For this reason, the four lowest categories were collapsed for multivariate analyses.

Figure A4: proportion localised by SES: (a) by period of diagnosis for M notified cases; (b) by notification method within period 3



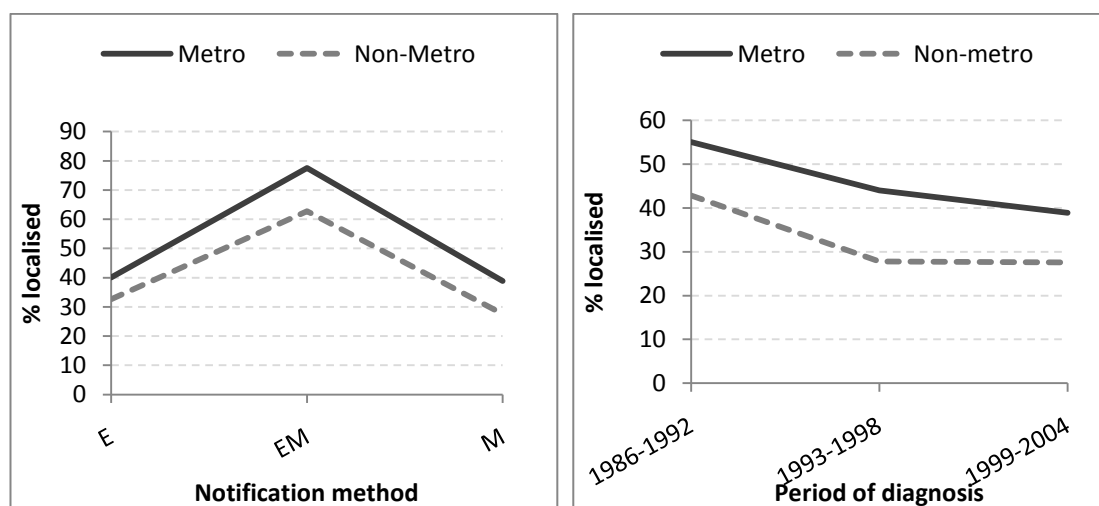
Area Health Service (AHS) of Residence at time of diagnosis

AHS at diagnosis showed a **significant association** with degree of spread ($\chi^2 = 383.1, p < 0.01$). The non-metropolitan areas (Hunter New England, North Coast, Greater Western and Greater Southern) had the highest proportion of unknown degree of spread cases. The areas were re-grouped for further analyses as metropolitan and non-metropolitan areas. Metropolitan areas showed consistently higher percentages of localised cases across period and notification methods.

Table A3: Distribution of Localised and Unknown cases, by Area Health Service of residence

AHS of residence		Localised	Unknown	Total (localised + unknown)	P value (χ^2)
South Western Sydney	<i>n</i>	1756	1355	3111	<0.01
	%	56.4	43.6		
South Eastern Sydney/ Illawarra	<i>n</i>	1531	1531	3062	
	%	50.0	50.0		
Western Sydney	<i>n</i>	1266	852	2118	
	%	59.8	40.2		
Northern Sydney/ Central Coast	<i>n</i>	1380	1122	2502	
	%	55.2	44.8		
Hunter / New England	<i>N</i>	1219	1271	2490	
	%	49.0	51.0		
North Coast	<i>N</i>	516	1022	1538	
	%	33.5	66.5		
Greater Southern	<i>N</i>	523	728	1251	
	%	41.8	58.2		
Greater Western	<i>N</i>	355	493	848	
	%	41.9	58.1		
Unknown	<i>N</i>	<5	-	<5	
	%	100	0	100	

Figure A5: proportion localised by AHS: (a) by period of diagnosis for M notified cases; (b) by notification method within period 3



Aboriginal/ Torres Strait Islander (ATSI) status

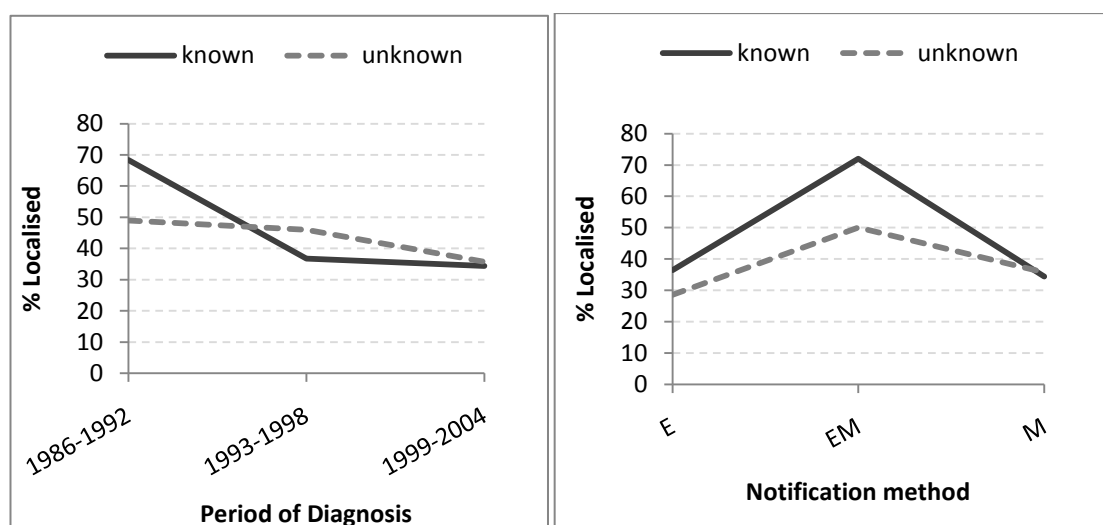
ATSI status is known to be under-reported in the CCR, but is recorded where sufficient information is available. “Unknown” ATSI status was treated as a separate category of interest. ATSI status **was significantly associated** with the proportion of unknown versus localised degree of spread cases ($\chi^2 = 49.08$, $p < 0.01$). Those persons of ATSI background had higher proportions of unknown degree of spread compared to those known to be non-ATSI. For those with unknown ATSI status there was an even higher proportion of unknown degree of spread cases. Due to low cell sizes, for further analyses, ATSI and Not ATSI were grouped as “known”.

There were distinct differences in the pattern of association across time periods and notification methods based on this re-grouping.

Table A4: Distribution of Localised and Unknown cases, by ATSI status

ATSI status		Localised	Unknown	Total (localised + unknown)	P value (χ^2)
ATSI	<i>n</i>	45	44	89	<0.01
	%	50.6	49.4		
Not ATSI	<i>n</i>	4481	3940	8421	
	%	53.2	46.8		
Unknown	<i>n</i>	4022	4390	8412	
	%	47.8	52.2		

Figure A6: proportion localised by ATSI status: (a) by period of diagnosis for M notified cases; (b) by notification method within period 3



Two Year Survival

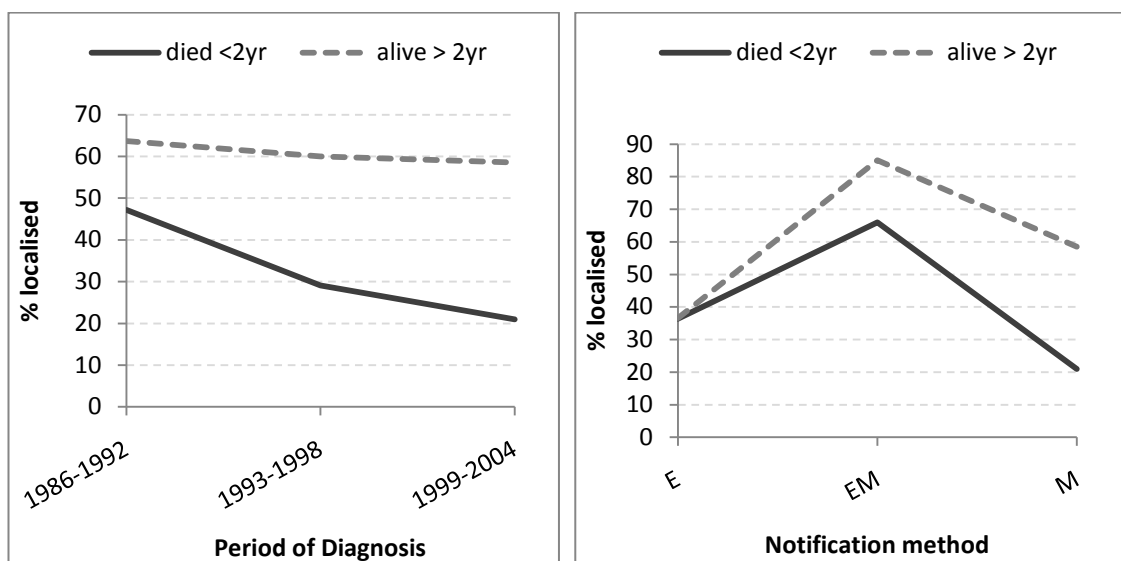
Two year survival **was significantly associated** with degree of spread with those surviving to 2 years more likely to have localised degree of spread compared to unknown degree of spread ($\chi^2 = 538.92, p < 0.01$).

Table A5: Distribution of Localised and Unknown cases, by two year survival status

Two Year Survival		Localised	Unknown	Total (localised + unknown)	P value (χ^2)
Did not survive to 2 years	<i>n</i>	5546	6764	12310	<0.01
	%	45.1	54.9		
Survived to 2 years	<i>n</i>	3002	1610	4612	
	%	65.1	34.9		

The association between two year survival and degree of spread was reasonably consistent across periods with the magnitude of effect increasing in later periods. However, the effect was inconsistent across notification methods, showing no effect within electronic only and a substantial effect within EM and M categories.

Figure A7: proportion localised by survival status: (a) by period of diagnosis for M notified cases; (b) by notification method within period 3



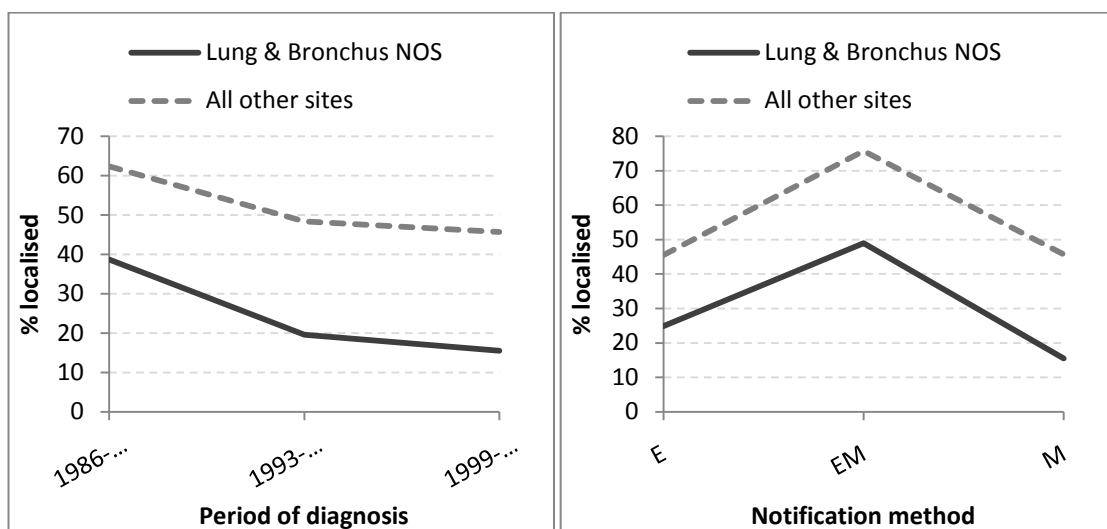
Site of Cancer

The site of the cancer within the lung **was significantly associated** with degree of spread ($\chi^2 = 1228.1, p < 0.01$). Cancers of the lung and bronchus that could not be otherwise specified (NOS) to a site were associated with much higher proportions of unknown compared to localised degree of spread. For further analyses, this NOS category was compared against a grouping of all other sites. This re-categorisation appeared to produce consistent effects across both diagnosis period and notification methods.

Table A6: Distribution of Localised and Unknown cases, by site of Lung cancer

Site of Cancer		Localised	Unknown	Total (localised + unknown)	P value (χ^2)
Trachea	<i>n</i>	27	20	47	<0.01
	%	57.4	42.6		
Main Bronchus	<i>n</i>	751	577	1328	
	%	56.6	43.4		
Upper Lobe	<i>n</i>	3464	2135	5599	
	%	61.9	38.1		
Middle Lobe	<i>n</i>	359	258	617	
	%	58.2	41.8		
Lower Lobe	<i>n</i>	1659	1004	2663	
	%	62.3	37.7		
Overlapping Lesion	<i>n</i>	83	39	122	
	%	68.0	32.0		
Lung and Bronchus NOS	<i>n</i>	2205	4341	6546	
	%	33.7	66.3		

Figure A8: proportion localised by site of cancer: (a) by period of diagnosis for M notified cases; (b) by notification method within period 3



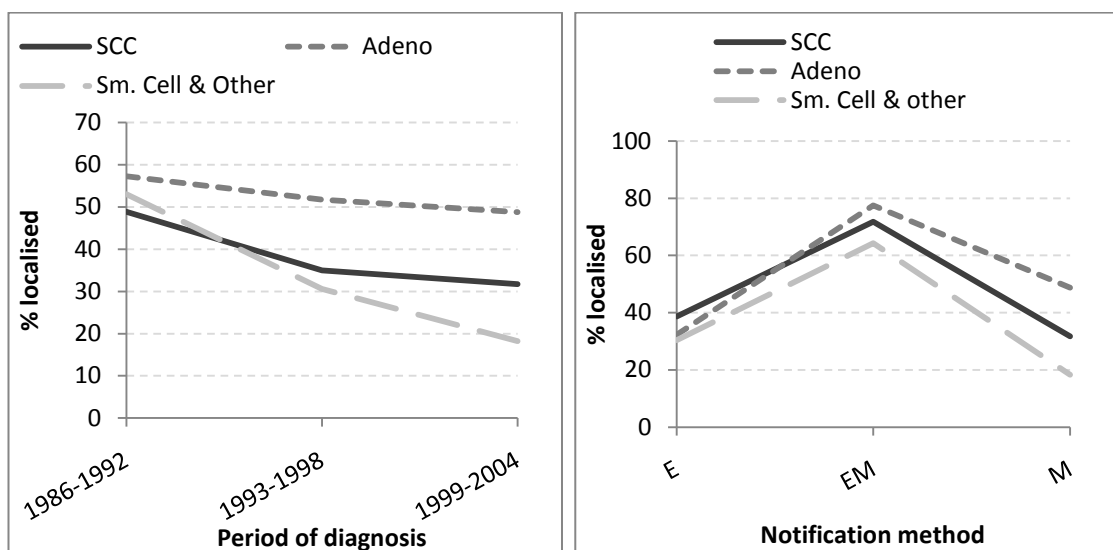
Histology

Cancers of the lung were grouped as Squamous Cell Carcinomas, Adenocarcinomas or Small-cell carcinoma. Histology group **was significantly associated** with degree of spread ($\chi^2 = 162.5, p < 0.01$). Squamous cell carcinomas were most likely to have unknown degree of spread with adenocarcinomas most likely to have localised degree of spread. The effect of histology did not appear to be consistent over time or notification method.

Table A7: Distribution of Localised and Unknown cases, by histology

		Localised	Unknown	Total (localised + unknown)	P value (χ^2)
Squamous cell carcinoma	<i>n</i>	5308	5588	10896	<0.01
	%	48.7	51.3		
Adenocarcinoma	<i>n</i>	2112	1518	3630	
	%	58.2	41.8		
Small-cell carcinoma	<i>n</i>	1023	1017	2040	
	%	50.1	49.9		
Other	<i>n</i>	105	251	356	
	%	29.5	70.5		

Figure A9: proportion localised by histology: (a) by period of diagnosis for M notified cases; (b) by notification method within period 3



Registration variables

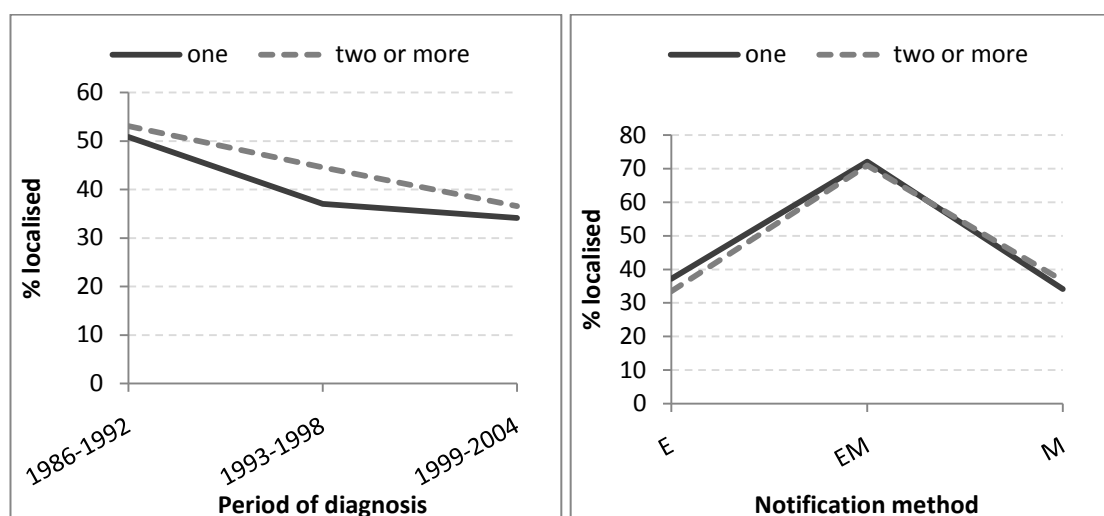
Number of primary cancers

The number of primary cancers per person **was not significantly** associated with unknown versus localised degree of spread ($\chi^2 = 0.17$, $p=0.68$). This variable showed slightly more association within period 2, but overall appeared to consistently show no or weak association with degree of spread across period and notification method.

Table A8: Distribution of Localised and Unknown cases, by number of primary cancers

Number of primary cancers		Localised	Unknown	Total (localised + unknown)	P value (χ^2)
1	<i>n</i>	7253	7086	14339	0.68
	%	50.6	49.4		
2 or more	<i>n</i>	1295	1288	2583	0.68
	%	50.1	49.9		

Figure A10: proportion localised by number of primary cancers: (a) by period of diagnosis for M notified cases; (b) by notification method within period 3



Method of diagnosis

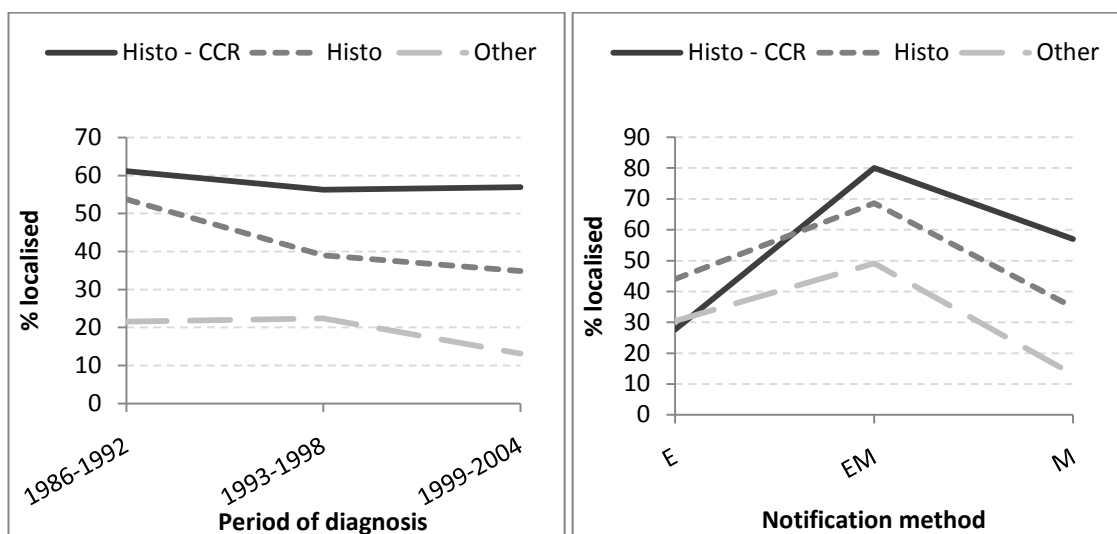
Method of diagnosis showed a **highly significant association** with degree of spread ($\chi^2 = 657.8, p < 0.01$). If diagnosis was based on methods such as cytology only, clinical notes, imaging or biochemistry only (grouped as other), it was more likely that degree of spread was coded as “unknown”. If histopathology was available and particularly if the histology was sighted by staff at the Central Cancer Registry then it was much less likely that a cancer was coded as unknown.

This variable demonstrated reasonably consistent effect across period but not across notification methods. Histology being sighted within the CCR was related to a higher degree of localised cancers for both EM and M notified cancers but not for E notified cases in period 3.

Table A9: Distribution of Localised and Unknown cases, by Method of diagnosis

Method of diagnosis		Localised	Unknown	Total (localised + unknown)	P value (χ^2)
Other	<i>n</i>	996	3247	4243	<0.01
	%	23.5	76.5		
Histopathology	<i>n</i>	2939	2610	5549	
	%	53.0	47.0		
Histo sighted at CCR	<i>n</i>	4613	2517	7130	
	%	64.7	35.3		

Figure A11: proportion localised by method of diagnosis: (a) by period of diagnosis for M notified cases; (b) by notification method within period 3



Number of Notification Episodes

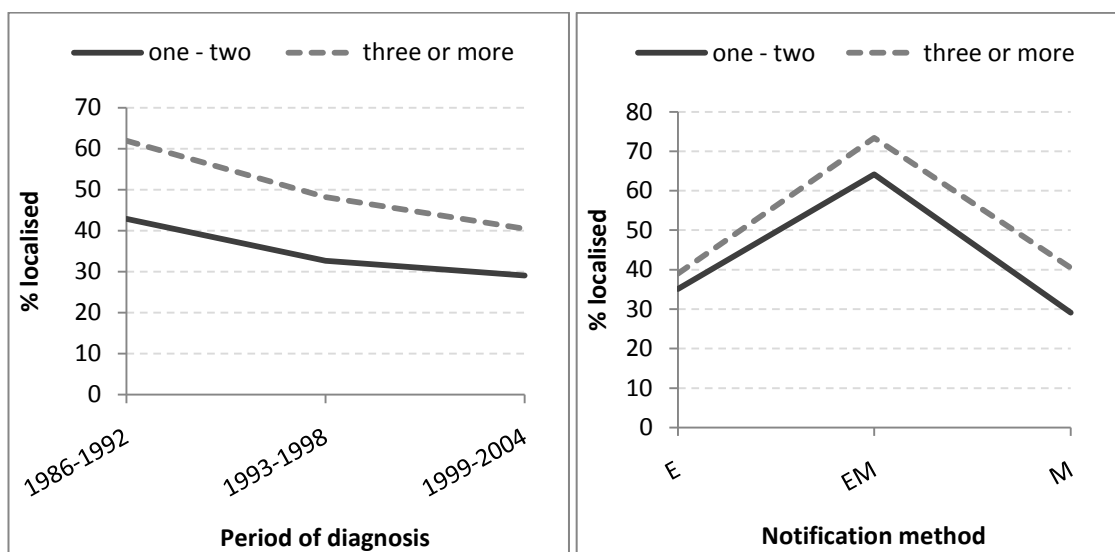
Number of notification episodes was categorised as “one-two” and “three or more”. Initial examination of association suggested that there was most difference between receiving only one notification versus receiving multiple notifications. However, the notification method grouping of electronic plus manual by definition requires at least two notification episodes, so maintaining a single notification category or treating this variable as continuous would be problematic. Based on this grouping, the number of notification episodes received **was significantly related** to degree of spread ($\chi^2 = 1828.3$, $p < 0.01$). Having received three or more notifications compared to one or two only was associated with a higher proportion of localised cases.

This variable behaved reasonably consistently across both period of diagnosis and notification method.

Table A10: Distribution of Localised and Unknown cases, by number of notification episodes

No. notification episodes		Localised	Unknown	Total (localised + unknown)	P value (χ^2)
One-Two	<i>n</i>	3304	4887	8191	
	%	40.3	59.7		
Three or more	<i>n</i>	5244	3487	8731	
	%	60.1	39.9		

Figure A12: proportion localised by number of notification episodes: (a) by period of diagnosis for M notified cases; (b) by notification method within period 3



Notification – Facility type

Multiple notifications could be received regarding a single cancer case and these could come from many different types of notifying institutions, including public hospitals, private hospitals, pathology labs and nursing homes. As there were many different combinations of responses possible, these were rationalised into three mutually exclusive groupings: those that included a private hospital notification; those that included a public hospital notification (but no private); and those that included neither a public or private hospital notification (classed as other).

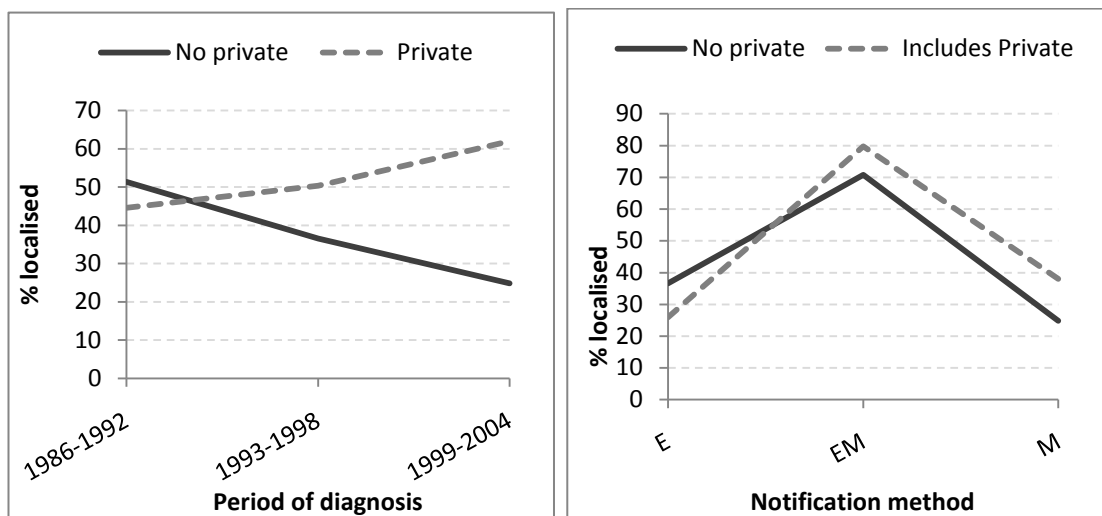
The type of facility type grouping submitting the notification(s) **was significantly related** to degree of spread ($\chi^2 = 577.3$, $p < 0.01$). If a notification was received from a private hospital, the degree of spread was most likely to be localised, with cases where no hospital (public or private) provided a notification much more likely to be “unknown”. Facility type was related to notification method with “other” notifications always submitted manually. For this reason, categories were re-grouped as “includes private” and “does not include private”.

This variable did not appear to behave consistently across period or notification method.

Table A11: Distribution of Localised and Unknown cases, by notifying facility type

Notification – Facility type		Localised	Unknown	Total (localised + unknown)	P value (χ^2)
Includes public but no private	<i>n</i>	7374	6936	14310	<0.01
	%	51.5	48.5		
Includes private	<i>n</i>	1032	629	1661	
	%	62.1	37.9		
Other (no public or private)	<i>n</i>	142	809	951	
	%	14.9	85.1		

Figure A13: proportion localised by types of notifying facility: (a) by period of diagnosis for M notified cases; (b) by notification method within period 3



Appendix B: Logistic Regression results for full model (model 1)

Table B1: Logistic Regression results for Model 1 applied to M notified cases in a) 1986-1992, b) 1993-1998 c) 1999-2004

		Model 1							
		M - 1986-1992		M - 1993-1998		M - 1999-2004		Consistent across period	
Parameter	Reference category	Estimate	Pr > ChiSq	Estimate	Pr > ChiSq	Estimate	Pr > ChiSq		
Intercept		-0.17	0.02	-0.28	0.00	-0.76	<.0001		
Age Group	65-74	<65	0.05	0.14	0.03	0.51	0.02	0.73	yes
	75+	<65	-0.14	0.00	-0.07	0.19	0.00	0.97	
Sex	F	M	-0.04	0.11	0.03	0.48	0.07	0.15	yes
SES	Other	Highest	-0.13	0.00	-0.17	0.00	0.03	0.66	no
Area of Residence	Non-metro	Metro	0.16	<.0001	0.26	<.0001	0.23	<.0001	yes
ATSI status	Known	Not known	0.15	0.00	-0.09	0.05	-0.02	0.73	no
Survival	alive > 2yr	Died<2yr	0.15	<.0001	0.39	<.0001	0.47	<.0001	yes
Histology	Adenocarcinoma	SCC	-0.26	<.0001	-0.38	<.0001	-0.43	<.0001	yes
	Sm. Cell and other	SCC	0.04	0.30	0.28	<.0001	0.34	<.0001	
Site	Bronchus & Lobes NOS	Other	-0.07	0.11	-0.29	0.00	-0.51	<.0001	yes
Number Primary Cancers	One	Two +	0.01	0.70	-0.01	0.75	0.07	0.21	yes
Method	Histo sighted at CCR	Other	0.46	<.0001	0.42	<.0001	0.69	<.0001	yes
	Histopathology	Other	0.37	<.0001	0.08	0.16	0.15	0.04	
Episodes	One-two	Three +	-0.16	<.0001	-0.11	0.00	0.11	0.03	no
Facility type	Includes private	No private	-0.20	0.00	0.15	0.00	0.59	<.0001	no
Adj r squared			0.18		0.25		0.39		
% Concordance			70.2		75.4		82.3		
% Discordance			29.2		24.2		17.5		
Goodness of fit			21.38	0.01	18.40	0.02	19.95	0.01	

TableB2: Logistic Regression results for Model 1 applied to a) M notified cases, b) EM notified cases and c) E notified cases in period 3

Model 1									
Parameter		Reference category	M - 1999-2004		EM - 1999-2004		E - 1999-2004		Consistent across notification method
			Estimate	Pr > ChiSq	Estimate	Pr > ChiSq	Estimate	Pr > ChiSq	
Intercept			-0.76	<.0001	0.01	0.98	-1.20	0.02	
Age Group	65-74	<65	0.02	0.73	0.02	0.75	0.03	0.70	yes
	75+	<65	0.00	0.97	0.03	0.60	0.13	0.10	
Sex	F	M	0.07	0.15	0.03	0.49	-0.01	0.86	no
SES	Other	Highest	0.03	0.66	-0.01	0.87	-0.17	0.05	no
Area of Residence	Non-metro	Metro	0.23	<.0001	0.30	<.0001	0.09	0.14	yes
ATSI status	Known	Not known	-0.02	0.73	0.65	0.05	0.06	0.90	yes
Survival	alive > 2yr	Died<2yr	0.47	<.0001	0.39	<.0001	-0.03	0.69	no
Histology	Adenocarcinoma	SCC	-0.43	<.0001	-0.35	<.0001	-0.45	<.0001	no
	Sm. Cell and other	SCC	0.34	<.0001	0.21	0.00	-0.11	0.37	
Site	Bronchus & Lobes NOS	Other	-0.51	<.0001	-0.26	0.00	-0.17	0.13	yes
Number Primary Cancers	One	Two +	0.07	0.21	0.09	0.10	0.11	0.12	yes
Method	Histo sighted at CCR	Other	0.69	<.0001	0.43	<.0001	-0.19	0.53	no
	Histopathology	Other	0.15	0.04	0.10	0.08	0.41	0.01	
Episodes	One-two	Three +	0.11	0.03	-0.10	0.08	0.01	0.92	no
Facility type	Includes private	No private	0.59	<.0001	0.14	0.03	-0.26	0.27	no
Adj r squared			0.39		0.16		0.11		
% Concordance			82.3		71.4		66.8		
% Discordance			17.5		28.1		32.5		
Goodness of fit			19.95	0.01	15.51	0.05	19.12	0.01	