COMPUTER METHODS AND PROGRAMS IN BIOMEDICINE 117 (2014) 412–424

journal homepage: www.intl.elsevierhealth.com/journals/cmpb

Automated estimation of disease recurrence in head and neck cancer using routine healthcare data

K. Ricketts[a], M. Williams[b,c,*], Z.-W. Liu[d], A. Gibson[a]

[a] Department of Medical Physics and Bioengineering, University College London, UK
[b] Radiotherapy Department, University College London Hospital, London, UK
[c] Department of Clinical Oncology, Imperial College Healthcare Trust, Charing Cross Hospital, Fulham Palace Road, London W6 8RF, UK
[d] ENT Department, Whipps Cross University Hospital, Whipps Cross Road, Leytonstone, London E11 1NR, UK

ARTICLE INFO

Article history:
Received 30 April 2014
Received in revised form
4 August 2014
Accepted 28 August 2014

Keywords:
Head and neck cancer
Disease recurrence
Routine datasets

ABSTRACT

Background: Overall survival (OS) and progression free survival (PFS) are key outcome measures for head and neck cancer as they reflect treatment efficacy, and have implications for patients and health services. The UK has recently developed a series of national cancer audits which aim to estimate survival and recurrence by relying on institutions manually submitting interval data on patient status, a labour-intensive method. However, nationally, data are routinely collected on hospital admissions, surgery, radiotherapy and chemotherapy. We have developed a technique to automate the interpretation of these routine datasets, allowing us to derive patterns of treatment in head and neck cancer patients from routinely acquired data.

Methods: We identified 122 patients with head and neck cancer and extracted treatment histories from hospital notes to provide a gold standard dataset. We obtained routinely collected local data on inpatient admission and procedures, chemotherapy and radiotherapy for these patients and analysed them with a computer algorithm which identified relevant time points and then calculated OS and PFS. We validated these by comparison with the gold standard dataset. The algorithm was then optimised to maximise correct identification of each timepoint, and minimise false identification of recurrence events.

Results: Of the 122 patients, 82% had locally advanced disease. OS was 88% at 1 year and 77% at 2 years and PFS was 75% and 66% at 1 and 2 years. 40 patients developed recurrent disease. Our automated method provided an estimated OS of 87% and 77% and PFS of 87% and 78% at 1 and 2 years; 98% and 82% of patients showed good agreement between the automated technique and Gold standard dataset of OS and PFS respectively (ratio of Gold standard to routine intervals of between 0.8 and 1.2). The automated technique correctly assigned recurrence in 101 out of 122 (83%) of the patients: 21 of the 40 patients with recurrent disease were correctly identified, 19 were too unwell to receive further treatment and were missed. Of the 82 patients who did not develop a recurrence, 77 were correctly identified and 2 were incorrectly identified as having recurrent disease when they did not.

* *Corresponding author at*: Department of Clinical Oncology, Imperial College Healthcare Trust, Charing Cross Hospital, Fulham Palace Road, London W6 8RF, UK.
E-mail address: matthew.williams2@imperial.nhs.uk (M. Williams) .

http://dx.doi.org/10.1016/j.cmpb.2014.08.008

Conclusions: We have demonstrated that our algorithm can be used to automate the interpretation of routine datasets to extract survival information for this sample of patients. It currently underestimates recurrence rates due to many patients not being well-enough to be treated for recurrent disease. With some further optimisation, this technique could be extended to a national level, providing a new approach to measuring outcomes on a larger scale than is currently possible. This could have implications for healthcare provision and policy for a range of different disease types.

1. Introduction

Measuring clinical outcomes in cancer patients remains challenging. Studies often report results from single centres, from large national studies, or from clinical trials. Single centre studies often provide detailed data, but in small numbers of treated patients, while national studies have more patients but lack details of treatment and outcomes. Clinical trials may provide high-quality data, but in small groups of highly selected patients, and are expensive to perform. Therefore the challenge remains to report high-quality, detailed clinical outcomes in large numbers of patients, without relying on specialised data collection.

Head and neck (H&N) cancers represent a large, heterogeneous group of cancers with ~460,000 cases worldwide [1]. Treatment often involves an intensive combination of surgery, radiotherapy and chemotherapy and patients can suffer significant long-term side effects. Despite this, tumour recurrence rates remain high and survival rates are relatively poor (*National Head & Neck Cancer audit 2011* [2]). Similar treatments may be given at recurrence, depending on the pattern of recurrence and the patient's ability to tolerate further treatment, but are toxic and expensive. Therefore, assessment of treatment efficacy, through measuring recurrence rates and recurrence-free survival, is important for patients, clinicians and health services.

The UK has had a dedicated national audit database for H&N cancer (DAHNO) since 2004 which requires manual data entry. As a consequence, by 2012 only 11.5% of patients had data on recurrence (*National Head & Neck Cancer audit 2011* [2]) Over this time period, the scope and scale of routinely-collected electronic healthcare data has expanded, often driven by the requirement for data for payment systems. In the UK, there are a variety of data sources, including data on hospital admissions and procedures (Hospital Episode Statistics; HES), radiotherapy (Radiotherapy Dataset; RTDS), chemotherapy (Systemic Anti-Cancer Therapy dataset; SACT) and deaths (Patient Demographics Service; PDS).

Learning how to use these data to answer clinically-driven questions, in a timely, accurate and relevant fashion represents a significant challenge [3]. A computerised method of accessing and interpreting such data can potentially yield useful clinical patterns and outcomes that are currently measured by national audits at considerable expense.

We have previously described a pilot study presenting a method of estimating recurrence and survival in H&N cancer patients based on manual analysis of routine data in a small group of 20 patients [4]. This work was based on capturing clinical intuitions about patterns of care and treatment in the form of simple rules about intervals between different treatment types. Since the routinely available data do not contain information on treatment intent, potentially curative treatments that were close to each other in time were assumed to be as part of a planned pattern of sequential, curative treatments, whereas if there was a significant gap between initial curative treatment and subsequent treatment, that subsequent treatment was assumed to be for recurrent disease. However, the assumptions underlying this approach are unlikely to be uniformly correct, and have never been validated in a large patient group.

In this study we:

1. Extend the range of clinical intuitions that we capture in computational form
2. Introduce a novel computer-based automated framework, written in Python, to merge and interpret the routine data and automatically identify diagnosis and recurrence events/dates;
3. Use that as a basis for performing simple in silico experiments to validate and optimise our approach in a larger patient sample ($n = 122$).

2. Method

2.1. Patient group

Patients from a single cancer centre (UCLH, London), were included if their first diagnosis of H&N cancer was made between 2009–2012, they had radiotherapy given with radical intent (either as primary modality or for recurrence), we were able to identify their date of diagnosis, they attended at least one follow-up visit at UCLH post-diagnosis, and they had an eligible histological subtype (squamous cell carcinoma, undifferentiated nasopharyngeal carcinoma, adenoid cystic carcinoma). The identification and exclusion of patients is shown in Fig. 1.

2.2. Creation and manual analysis of gold standard dataset

We manually extracted data on patient characteristics, staging and treatment from hospital records. The date of diagnosis was the date that the first diagnostic specimen was obtained.

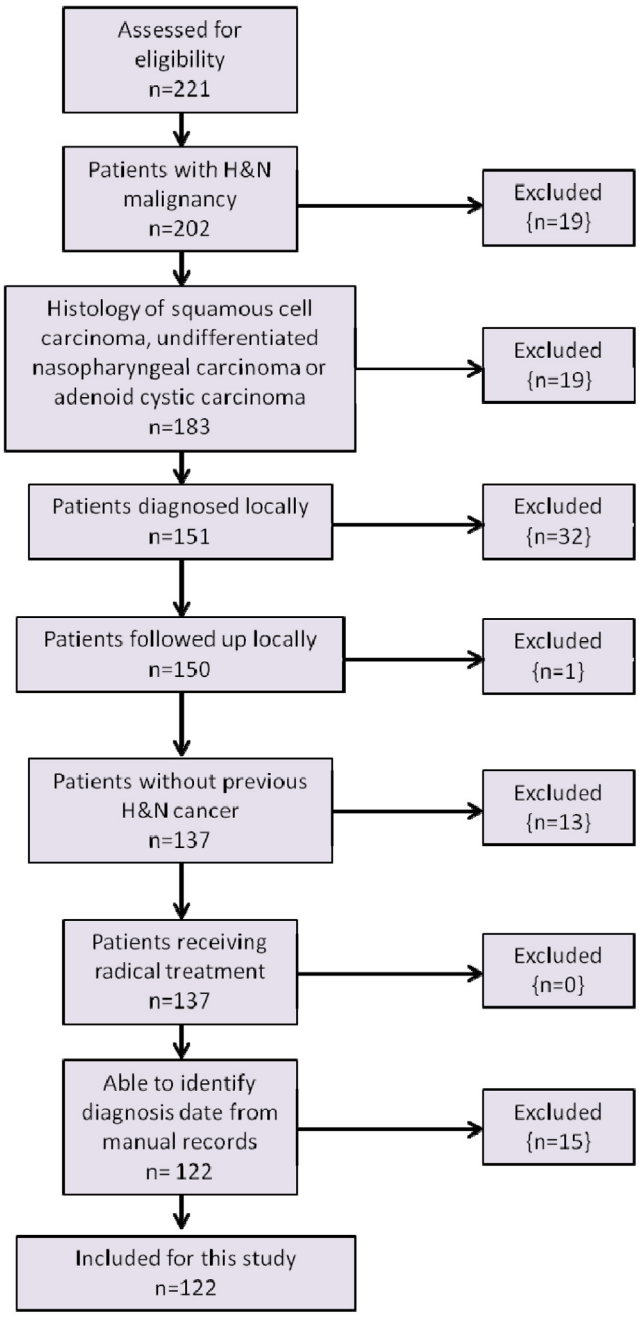

Fig. 1 – Patient identification and selection flowchart.

The date of recurrence was the earliest date that recurrence was confirmed, on clinical, radiological or histological grounds. The date of last follow-up was the latest date of known contact with the hospital. The overall survival interval was the time between diagnosis and death or last follow-up. The progression-free survival interval was the time between diagnosis and evidence of recurrent or progressive disease (see Fig. 2, upper half). Patients who were lost to follow-up, or were alive at the conclusion of the study were censored at the time of last known follow-up (according to the Kaplan–Meier method). This provided a "Gold standard" dataset of manually curated and analysed data.

2.3. *Routine dataset*

We obtained local data sources for each patient from the Hospital Information Department, UCLH. These were:

1. Personal demographic service (PDS): notification of death, and date of death if dead
2. Hospital episode statistics (HES): records of start and end dates of admissions and inpatient procedures and diagnoses
3. Chemotherapy data (SACT): records of administration of chemotherapy and diagnoses and dates of treatments
4. Radiotherapy data (RTDS): records of delivery of radiotherapy and diagnoses and dates of treatments.

All data sources use the International Classification of Diseases v 10 (ICD-10) for diagnostic information. HES uses the Office of Population and Census Classification of Interventions and Procedures v 4 (OPCS-4). ICD-10 is widely used internationally, and both have mappings to other terminologies such as SNOMED-CT.

Our algorithm integrated the extracted HES, SACT, RTDS and PDS data to form a single list of event data for each patient, comprising diagnostic (ICD-10) and interventional codes (OPCS 4.4). The events were arranged in chronological order and filtered to ensure only treatments relevant to head and neck cancer were analysed (according to ICD10 codes relating to malignant neoplasms of the head, face or neck as displayed in Appendix 1.1). This information was used to inform the development of our algorithm.

2.4. *Automated interpretation of routine data*

None of the routine data sources (RTDS, SACT, HES and PDS) directly report either date of diagnosis or date of disease recurrence, and so we defined proxy time points for relevant clinical events. Automated strategies were used to identify these proxy time points and thus estimate survival.

Date of diagnosis was taken to be the date of the first recorded ICD-10 code in HES that corresponded to a diagnosis of head and neck malignancy (see Appendix 1.1).

Date of recurrence. We used the start date of secondary treatment for HNSCC (whether surgery, radiotherapy or chemotherapy) as a proxy for recurrence. We initially defined a treatment as being for a recurrence if it was given more than 90 days after the end of the previous treatment if two treatments occurred within this time interval they were considered to form a planned primary treatment strategy, otherwise the later treatment was assumed to be treatment for recurrent disease.

All routine data (HES, SACT, RTDS and PDS) were extracted on a per-patient basis. Blank entries were removed. Radiotherapy and chemotherapy data were summarised and ordered chronologically, and used as the basis for estimating OS and PFS. For the purposes of this study, we compared the output of software with the manually curated gold standard dataset to allow us to assess the accuracy of the automated approach.

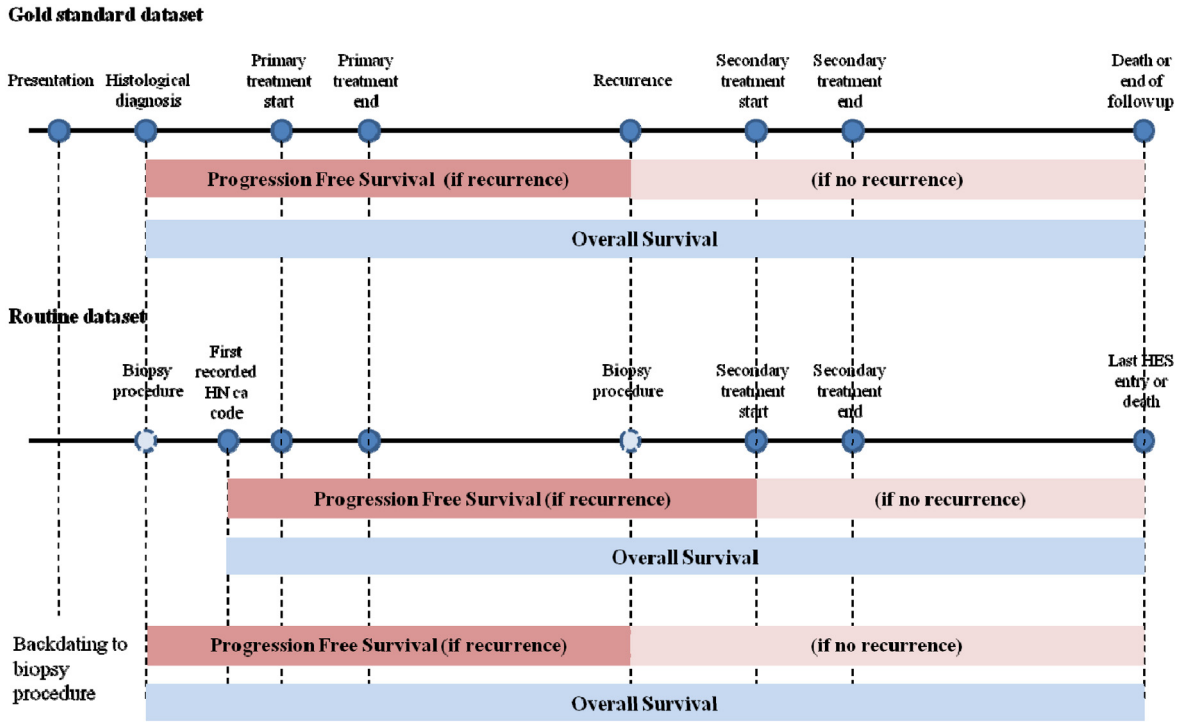

Fig. 2 – Definition of progression free survival and overall survival for the Gold standard and routine datasets.

2.5. *Optimisation of automated method*

We further optimised our approach through use of a time-based threshold to distinguish between a series of primary treatments, which may follow on from one another, and treatments given in response to disease recurrence. Our manual pilot study [4] used a static time interval of 90 days between consecutive treatments to distinguish delayed primary treatment from treatment for recurrence. In this work, we explored systematic optimisation of this time interval via in silico experimentation. Since both diagnosis and recurrence are often preceded by coded diagnostic procedures, we investigated whether backdating diagnosis and recurrence would improve estimated OS and PFS. We backdated the date of diagnosis or recurrence to the earliest of:

1. The date of the earliest diagnostic procedure within a set time interval of a treatment for H&N cancer; this time interval was initially set at 42 days, and then varied to find the optimal interval.
2. The first date of metastatic disease (determined by the appearance of a metastatic cancer ICD-10 code in HES, appendix 1.1).
3. The start of radiotherapy for an H&N cancer.

Since a longer backdating time interval risks inclusion of irrelevant diagnostic procedures, we attempted to find the shortest time interval (used for backdating for both diagnosis and recurrence) consistent with the optimal agreement with the gold standard data. For each experiment, overall survival and progression-free survival intervals were calculated using the proxy diagnosis and recurrence time points as illustrated in Fig. 2 (lower half). We assessed the impact of our experiments based on the agreement between the automated results and the gold standard data in three areas:

1. Recurrence events (assessed by calculating the proportion of patients in whom the recurrence was correctly detected by the automated data vs. the gold standard).
2. Overall survival interval (calculated on an individual patient basis); we found the survival intervals in the gold standard and the routine data set, and took the ratio. In line with other work [5,6] a ratio of 0.8–1.2 was taken to represent reasonable agreement.
3. Progression-free survival interval (calculated individually and presented as a ratio, as for overall survival time).

Since we could not assume that correlation was normally distributed we used Kendall's tau to assess correlation using the statistical software, R [7].

The routine data was supplied as a set of text (CSV) files. We wrote our own software (in Python v3.2) to integrate data from the different data sources, to automatically identify diagnosis and recurrence time points, to calculate survival intervals, and to determine the agreement between the automatic analysis and the gold standard dataset. These results were then exported as a text file for subsequent analysis in the open-source statistics software, R. Fig. 3 shows the flow of data through the automated software.

The software requires five sets of input data:

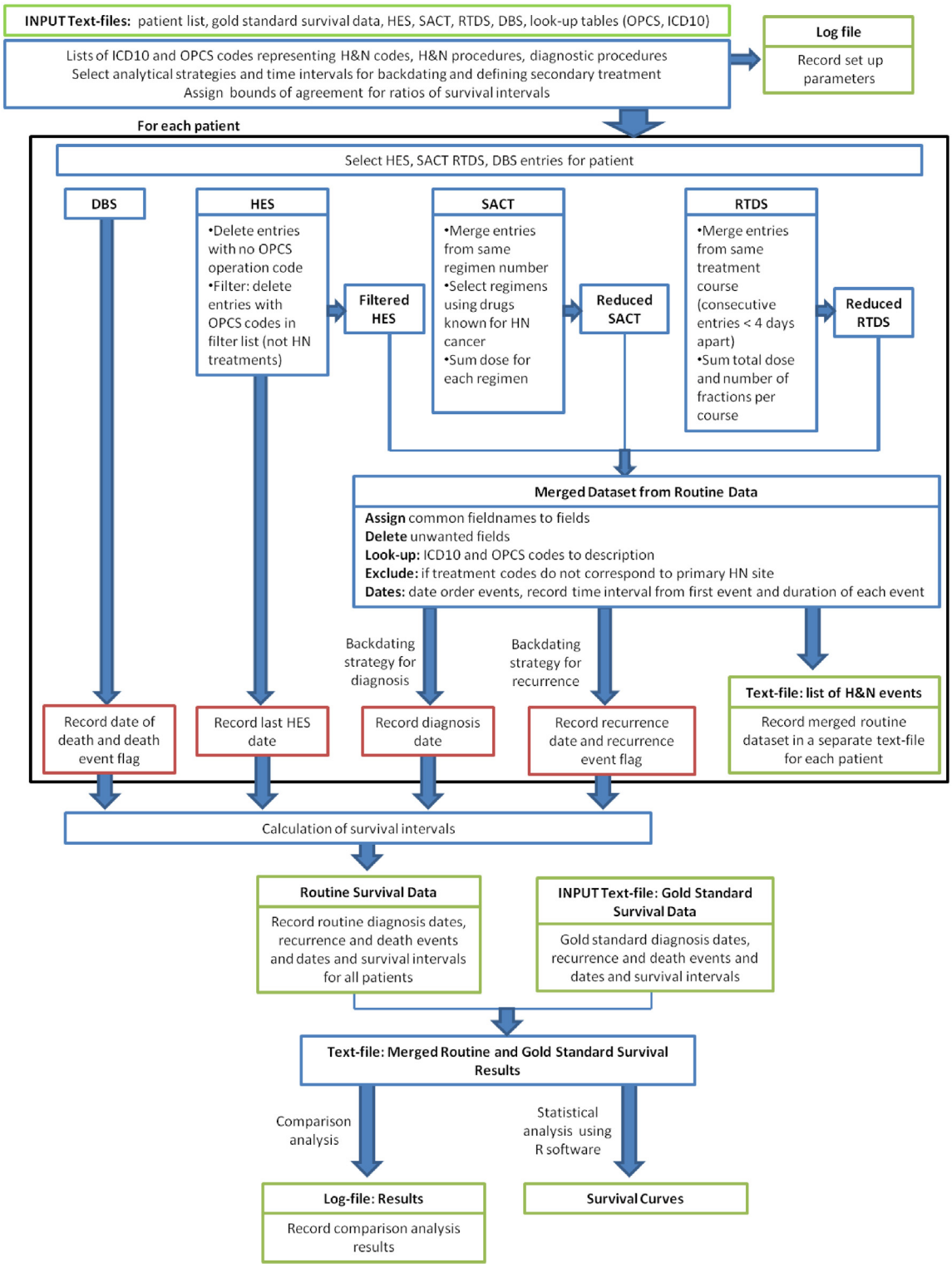

Fig. 3 – Flow chart of python software for the automated identification of diagnostic and recurrence time points from routine data. The section in the lower half of the figure, referring to the integration of routine and gold-standard data, was to allow us to assess the accuracy of the routine-data based approach.

- A list of anonymised patient identifiers, used to allow us to link data from different sources
- Routine data sources (from HES, SACT, RTDS and PDS)
- Look-up tables to translate ICD10 and OPCS codes to text (Appendices 1.1 and 1.2)
- A list of procedures that are not considered signs of recurrence (Appendix 1.3)

Table 1 – Staging data for patients included in the study.

Tumour site	Total	Stage I	Stage II	Stage III	IVa	IVb	IVc	Unknown
Oropharynx	51	0	1	5	37	2	1	5
Larynx	26	10	2	4	9	0	0	1
Hypopharynx	13	0	0	1	9	1	1	1
Oral cavity	13	1	0	0	10	0	0	2
Other sites or primary unknown	19	0	1	3	6	1	0	8
Total	122	11	4	13	71	4	2	17

- A list of diagnostic procedures (e.g. biopsy: Appendix 1.4).

3. Results

122 patients met the inclusion criteria (Fig. 1). The median age was 61 (range 20–82) and 82% of the patients presented with locally advanced disease (stages III–IV). The tumour sites and staging data are presented in Table 1. 40 of the patients developed recurrent or progressive disease. Overall survival was 88% at 1 year and 77% at 2 years. Progression-free survival was 75% and 66% at 1 and 2 years. Our initial, automated, unoptimised method provided: an estimated OS of 87% and 77% and PFS of 87% and 78%: 21 of the 40 patients with recurrent disease were correctly identified and 19 were missed; of the 82 patients who did not develop a recurrence, 77 were correctly identified and 5 were incorrectly identified.

3.1. *Optimisation of automated method*

3.1.1. *Optimisation by backdating strategy*

We investigated the impact of backdating the date of diagnosis and recurrence to the date of biopsy, radiotherapy treatment or date of presentation with metastatic disease. The results of this are summarised in Table 2. Backdating to the date of biopsy improved the number of patients in whom the dates of diagnosis and recurrence were in agreement with gold-standard data, as did backdating to the start of radiotherapy. Backdating to a diagnosis of metastatic disease had no impact, and we therefore disregarded it in further analysis. Combining backdating to biopsy or start of radiotherapy resulted in a modest improvement in performance above our initial technique, reducing the number of patients where there was significant error (ratio of automated/gold standard outside 0.8–1.2) in OS and PFS interval by 57% and 8% respectively. However, none of the backdating strategies improved the identification of disease recurrence.

3.1.2. *Optimisation of time intervals for backdating biopsy*

Since backdating to biopsy and start of radiotherapy by up to 6 weeks improved performance modestly, we systematically varied the interval from 14 to 70 days, in 14 day increments. On the basis of this, a time interval of 56 days was the shortest interval with the best agreement with the gold standard data for both diagnosis and recurrence (Figs. 4 and 5). Altering the time intervals used for backdating improved the agreement of dates further, but did not improve the number of patients in whom we detected recurrence.

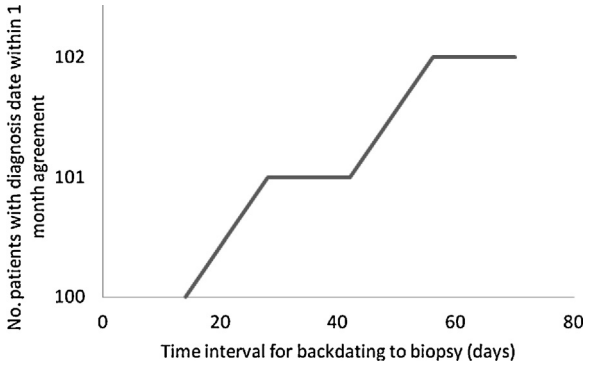

Fig. 4 – Effect of modifying time interval for backdating diagnosis date to biopsy procedure on the number of patients with estimated date within one month of Gold standard date. No improvement was seen for diagnosis dates in perfect agreement or within 1 week's agreement upon modifying the biopsy backdating time interval.

3.1.3. *Use of time interval after end of primary treatment to define secondary treatment*

The results of optimal definition of the interval between one treatment and another to distinguish between planned adjuvant treatment and treatment for recurrent disease are displayed in Table 3. We aimed to minimise the incorrect classification of adjuvant treatment and to maximise correct identification of treatment for recurrent disease. A time interval of 120 days resulted in the fewest recurrence event disagreements (21) and falsely identified recurrence events ($n = 2$ events). This results in an optimal specificity of 97.6% and sensitivity of 52.5% for detecting recurrence events.

Using a 120-day interval to differentiate between planned adjuvant treatment and treatment of recurrent disease also gave the best agreement between the gold standard and automated PFS intervals (Fig. 6). Since this did not increase

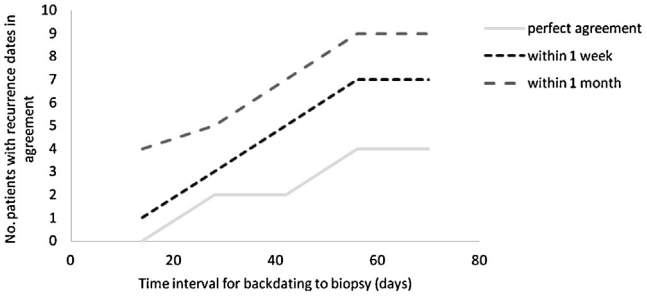

Fig. 5 – Effect of modifying time interval for backdating recurrence date to biopsy procedure.

Table 2 – Impact of backdating diagnosis and recurrence dates to (A), biopsy (backdating time interval t = 6 weeks); (B) start of radiotherapy (no time interval condition imposed); and (C) metastatic ICD10 diagnosis codes (t = 6 weeks). A secondary treatment definition time interval of >3 months was used in each case.

Experimental Condition	No. patients out of bounds for routine OS	No. patients out of bounds for routine PFS	Diagnosis dates in agreement {$n = 122$} ± 1 week/ ± 1 month	Recurrence dates in agreement {$n = 40$} ± 1 week/ ± 1 month	No. of recurrence events correctly identified	No. of false positive recurrence events	No of missed recurrence events
No backdating	7	25	1 week (62) 1 month (97)	1 week (1) 1 month (4)	21	5	19
A	7	26	1 week (61) 1 month (98)	1 week (4) 1 month (6)	21	5	19
B	3	23	1 week (62) 1 month (100)	1 week (1) 1 month (4)	21	5	19
C	7	26	1 week (62) 1 month (97)	1 week (1) 1 month (4)	21	5	19

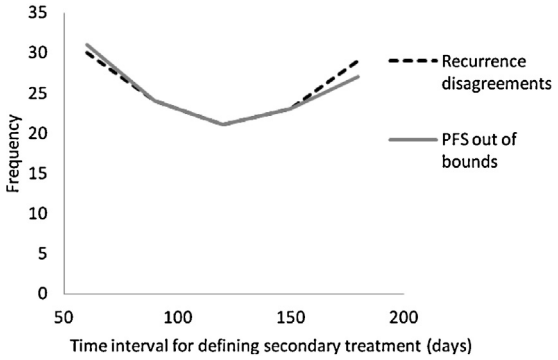

Fig. 6 – Effect of modifying time interval for defining secondary treatment on the number of patients for whom there was disagreement as to recurrent disease status between the Gold standard and routine datasets, and the number of patients with PFS intervals which were significantly different to those in the Gold standard dataset.

the number of recurrences missed and minimised the number incorrectly ascribed (see Table 3), this appears to be the optimal time-period in this group of patients.

3.1.4. *Summary*

The best agreement between the automated analysis and manual analysis of the gold standard data came from:

1. Backdating to a diagnostic event to no more than 56 days prior to the first evidence of a H&N cancer diagnosis code.
2. Backdating to radiotherapy with no time limit.
3. Assuming that treatment was for recurrent, rather than initial, disease if it occurred more than 120 days following the primary diagnosis.

Table 4 displays the results of the final optimised automated strategy.

We further stratified the results to see the influence on tumour type on successful identification of recurrence; results displayed in Table 5. We successfully predicted 78% of recurrences in laryngeal cancer, 54% in oropharynx- and 67% in oral cavity cancers. This is probably as laryngeal cancer tends to present at an earlier stage (and in our sample was the subsite with the largest number of patients with stages I and II disease) and therefore recurrences would be amenable to further radical treatment. This suggests that even though most of the patients in our sample had advanced disease, our method would be relatively robust if more early stage cancers were included.

3.2. *Survival and recurrence data*

There was good correlation for OS and PFS between the two datasets as displayed in Fig. 7. For OS, Kendall's tau was 0.97 (p-value < 0.0001), and for PFS, tau = 0.82 (p-value < 0.0001).

Table 3 – Number of recurrence date and event disagreements (the recurrence status of the patient is different in the gold standard and automated data), correctly identified (automated and gold standard both identify recurrence), falsely identified (automated data contains a recurrence not present in the gold standard data) and missed (automated data does not contain a recurrence that is present in the gold standard).

Time interval (days)	Recurrence dates in agreement $\{n=40\} \pm 1$ week/± 1 month	No. of recurrence events correctly identified	No. of recurrence events falsely identified	No. of recurrence events missed	Sensitivity	Specificity
60	1 week (6) 1 month (8)	21	11	19	52.5%	86.6%
90	1 week (7) 1 month (9)	21	5	19	52.5%	93.9%
120	1 week (7) 1 month (9)	21	2	19	52.5%	97.6%
150	1 week (6) 1 month (8)	19	2	21	47.5%	97.6%
180	1 week (4) 1 month (5)	13	2	27	32.5%	97.6%

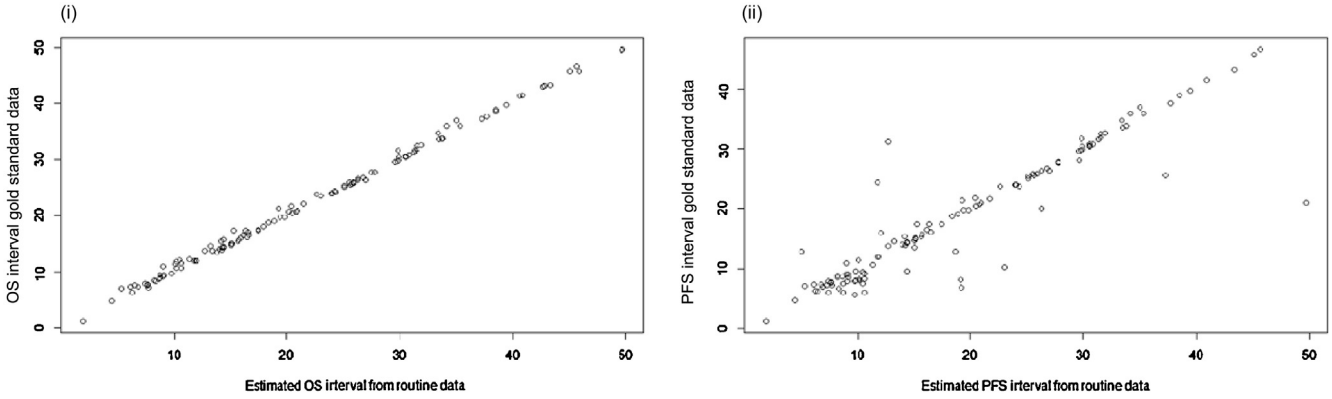

Fig. 7 – (i) Overall survival and (ii) progression free survival intervals of the automated routine data plotted against Gold standard data.

3.3. *Comparison of optimised automated technique and gold standard*

98% and 82% of patients showed good agreement between the automated technique and Gold standard dataset of OS and PFS respectively (ratio of Gold standard to routine intervals of between 0.8 and 1.2). 61 diagnosis dates out of 122 were correct to the nearest week of the true date. The automated technique correctly assigned recurrence in 101 out of 122 (83%) of the patients. Of the 40 patients who developed a recurrence, 21 were identified by the automated technique. Of these, four dates were perfectly identified, seven were within 1 week and nine were within 1 month of the Gold standard dataset. The automated technique failed to detect 19 patients who developed a recurrence, and incorrectly inferred that two patients had developed a recurrence when they had not.

4. Discussion

We have automated and optimised a novel computer-based approach for estimating disease-recurrence in head and neck cancer through a series of experiments. An earlier pilot study suggested that our approach may have some validity. This study has confirmed its feasibility in a larger group of patients using an automated approach, and has documented the specific impact of a range of automatic backdating strategies on the performance of our approach. The final automated approach successfully identified recurrence status in 83% of patients, and 98% and 82% of patients showed good agreement between the automated technique and gold standard dataset of OS and PFS respectively (ratio of gold standard to routine intervals of between 0.8 and 1.2).

Our optimisation procedure showed small yet measurable improvements in estimating OS and PFS intervals over the unoptimised technique. The diagnosis date was generally later than that for the gold standard by a median of 2 days, and the estimated OS interval was more than 20% different from the gold standard OS interval for only three patients. Backdating to biopsy was found to improve diagnosis date agreement; without backdating to biopsy the routine diagnosis date preceded that for the Gold standard by a median of 5 days.

However, backdating did not improve the proportion of patients who developed recurrent disease but were not detected by our approach. Of the 19 patients with recurrent disease who were missed, 16 (79%) of these failed cases were due to the patient not receiving treatment for their recurrence. One other was because the patient suffered from early recurrence 41 days after primary treatment and so a secondary treatment was assumed, following our automated technique,

Table 4 – Final performance of the approach, based on backdating of diagnosis and recurrence date with a 56 day interval, backdating to start of radiotherapy, and using a 120 day time interval for definition of secondary treatment.

Conditions	No. patients out of bounds for routine OS	No. patients out of bounds for routine PFS	Diagnosis dates in agreement $\{n = 122\} \pm 1$ week/± 1 month	Recurrence dates in agreement $\{n = 40\} \pm 1$ week/± 1 month	No. of recurrence events correctly identified	No. of recurrence events falsely identified	No. of recurrence events missed
Initial approach	7	25	1 week (62) 1 month (97)	1 week (1) 1 month (4)	21	5	19
Backdating alone	3	23	1 week (61) 1 month (101)	1 week (5) 1 month (7)	21	5	19
Backdating + optimised time intervals	3	21	1 week (61) 1 month (102)	1 week (7) 1 month (9)	21	2	19

to be planned primary treatment. Another missed recurrence had an incorrect ICD10 code, and the final missed recurrence was because the surgical treatment of recurrence was miscoded as a reconstructive procedure. Two recurrences were falsely detected in the routine data due to: (i) a wrong diagnosis codes (HN cancer rather than lung cancer) being assigned a radiotherapy treatment, and (ii) a patient receiving treatment coded as chemotherapy one year after primary radiotherapy, which was therefore used as an indication of recurrence by our automated approach. These results are summarised in Table 5.

There are several limitations to this work. Firstly, we have only looked at patients from one cancer centre in the UK, and we have concentrated our work on patients at higher risk of recurrence, rather than using a random sample. It will be important to assess the performance of our system in an unselected population, particularly in those with a lower risk of recurrence. However, we chose to use data which are available at a national level in the UK, and are based on widely used standards (OPCS, ICD-10), so that that they could be reasonably easily replicated. In particular, we have not used clinically important data which are not currently reliably collected in routine care, such as cancer stage. In addition, we accept that our results may not generalise to other populations. Our technique has moderate performance. It is not clear what level or performance is necessary for the technique to be clinically applicable, and previous studies have not examined performance in detail. We suspect that even having broad estimates may be useful, but there are also clear areas for development which are the focus of on-going work.

Aside from the specific problems in this dataset, there are two main problems in using routine health-care data for clinical purposes. The first is that such data, often collected by staff who are not involved in direct clinical care, may not accurately reflect the clinical situation. In our dataset, these were the causes for three of the errors. In theory, these could be reduced by training and increasing clinical involvement (see [8] for a discussion). The second source of error is in the mismatch which occurs in using administrative data for clinical purposes. These data are collected to inform payment and billing operations, rather than clinical care, and using them for clinical purposes often requires some degree of inference. In systematic terms, the development of progressive/recurrent disease is not directly captured in the routine data, and thus we infer recurrence from diagnostic and treatment activity that happens as a result. In this dataset, there was a specific instance where a patient received a high-cost drug regimen in the intensive care unit, which was coded as “chemotherapy”. The degree to which these inferences can be generalised and extended remains unclear, and is to some extent probably disease and health-care system specific.

Previous work in this area is extremely limited. Other authors have shown that routinely collected data can be used to estimate mortality rates, related to either different oncological treatments [9] or following orthopaedic surgical procedures [10]. There has also been work showing that cancer registry data can be used to measure recurrence rates in breast cancer [11]. However, there is very little work on using routinely available procedure-level data to infer disease recurrence. One study examined the use of such data to estimate measures of metastatic disease in breast, prostate and lung

Table 5a – Recurrence events correctly predicted and missed for each tumour site.

Tumour site	Total	Recurrences not predicted	Recurrence correctly predicted	% recurrence correctly predicted
Oropharynx	51	6	7	54
Larynx	26	2	7	78
Hypopharynx	13	6	1	14
Oral cavity	13	1	2	67
Other sites or primary unknown	19	4	4	50
Total	122	19	21	53

Table 5b – Numbers and causes of the failure of our automated approach.

Failure	Number of patients $\{n = 122\}$	Reason for failure
Recurrence falsely identified	2	•Incorrect diagnosis ICD10 code assigned (HN cancer instead of lung cancer) $\{n = 1\}$ •Primary chemotherapy treatment received 1 year after primary radiotherapy–wrongly detected as treatment for recurrence $\{n = 1\}$
Recurrence event missed	19	•No treatment for recurrence $\{n = 16\}$ •Incorrect diagnosis ICD10 code assigned (disorders of nose J348) $\{n = 1\}$ •Early recurrence (41 days after primary treatment) $\{n = 1\}$ •Treatment surgery miscoded as reconstructive procedure $\{n = 1\}$

cancer [12], and used a combination of ICD-9 codes for diseases and treatment codes for chemotherapy. The other study used patients enrolled in a clinical trial, and looked at their routine health claims data to estimate recurrence and death [13]. They found a good rate of agreement, with the routine healthcare data being able to measure the 5-year disease-free survival rate with a substantial degree of accuracy. In both cases, the authors used a combination of clinical intuition and logically-described criteria to use the routine data to infer recurrence. This approach is similar in principle to our work, although we included a wider range of treatment modalities than either of their studies did, and is the only one to include patients with head and neck cancers.

Our work has focused on head and neck cancer, where local recurrence is relatively common and rarely left untreated. However, we believe that our approach is applicable to other tumour sites, and possibly non-malignant diseases. Although there are some disease-specific aspects to the work (for example, the presentation of disease with initial lymph-node involvement in the neck), the principle of retrospectively assigning a date of diagnosis to a primary tumour based on the (earlier) date of presentation of metastatic disease would seem to be applicable across multiple tumour sites. Similarly, the use of repeated treatments to measure recurrent disease is one that is potentially applicable to a wide variety of conditions where recurrence is a possibility. However, it is restricted to those patients who are able or willing to receive subsequent treatments, and there are a proportion of patients who are either unwilling or unable to undergo further treatment, and thus will not be detected by our approach. Currently, our approach offers reasonable performance in a group of patients who are well enough to undergo initial curative treatment. There remain further possibilities for optimisation, through a more individualised assessment of relapse, and cross-referencing data on overall survival and local recurrence. These, and other developments, remain the subject of further work.

5. Conclusions

We have developed a computer-based automated technique to extract data on relevant clinical events from routine healthcare datasets (HES, SACT, RTDS and PDS). Diagnosis, recurrence and death dates, and date of last follow-up were identified and overall survival and progression free survival intervals were calculated from this data. The automated algorithm was optimised to maximise correct identification of each timepoint, and minimise false identification of recurrence events. We tested our algorithm on 122 patients who had received radical treatment for head and neck cancer; the recurrence status for 82% patients was correctly identified, and in 98% of patients there was acceptable agreement between the routine and gold standard dataset for overall survival intervals (83% for progression free survival). 21 recurrence events were correctly identified out of a total of 40. The 19 who were missed were mainly due to the patient not receiving treatment for their recurrence; secondary treatment was used as an indicator for recurrence in the algorithm. We have demonstrated that our algorithm can be used to automate interpretation of routine datasets to extract survival information for this sample of patients, and the coding schemes and approach that the technique uses opens it up to use at a national level. There is potential to develop this algorithm to sensitise the recurrence dating strategy to contraindications of the patient, and also to perform retrospective analysis by predicting likelihood of recurrence through knowing the final status of the patient.

Conflict of interest

No authors have been paid for this work, and all declare there are no conflicts of interest.

Acknowledgements

The authors are grateful to Primrose Carr and Martyn Turner for their informatics assistance, and to Maria Kilkenny for her insight into the head and neck patient pathway. We would like to thank Meredy Pritchard and Atia Khan for assistance in collating patients for the study, and to the H&N cancer team (both surgical and oncological) at UCH for their interest and support. KR and AG were funded by EPSRC grant EP/F01208X/1.

Appendix.

See Table A.1.1.
See Table A.1.2.
See Table A.1.3.
See Table A.1.4.

Table A.1.1 – List of ICD10 codes providing evidence of head and neck malignancy and evidence of metastasis.

Head and neck ICD10 codes	Description
C00–C14	Malignant neoplasms of lip, oral cavity and pharynx
C30–C32	Malignant neoplasms of nasal cavity, middle ear, sinuses and larynx
C76.0	Malignant neoplasms of ill-defined, secondary and unspecified sites–head, face and neck
C77.0	Secondary and unspecified malignant neoplasm of lymph nodes of head, face and neck
Metastasis ICD10 Codes	
C77.0	Secondary and unspecified malignant neoplasm of lymph nodes
C78.0	Secondary malignant neoplasm of respiratory and digestive organs
C79.0	Secondary malignant neoplasm of other sites

Table A.1.2 – List of OPCS 4.4 codes taken as evidence of recurrence if given as second modality treatment.

OPCS 4.4 codes	Description
E032	Excision of lesion of septum of nose
E191, E192	Total pharyngectomy, partial pharyngectomy
E291–E294, E296	Total laryngectomy, partial laryngectomy, laryngectomy NEC
E356	Endoscopic partial laryngectomy
E342, E343	Endoscopic resection/destruction of lesion of larynx
E352, E353	
F021, F022	Excision/destruction of lesion of lip
F221, F222, F228, F229	Total glossectomy, partial glossectomy, hemiglossectomy, glossectomy NEC unspecified
F231, F232	Excision/destruction of lesion of tongue
F341, F342–F349, F361	Bilateral dissection tonsillectomy, guillotine tonsillectomy, laser excision/coblation/unspecified/destruction of tonsil
G422, G428, G429, G431–G435, G438, G439	Photodynamic therapy of lesion of upper GI tract, extirpation/snare resection/laser destruction/cauterisation/sclerotherapy/destruction of lesion of upper GI tract
T851	Block dissection of cervical lymph nodes
V061, V068, V069, V071-V074, V078, V079	Maxillectomy, medial maxillectomy, other specified excision of maxilla, excision of bone of face
V141–V144, V148, V149	Hemimandibulectomy, extensive excision of mandible, mandibulectomy, partial excision of mandible
X654, X658, X659	Delivery of radiotherapy
X670, X671, X675, X673–X679	Preparation for radiotherapy
X701–X705, X708, X709, X711–X715, X718, X719	Procurement of drugs for chemotherapy
X721–X724, X728, X729, X731, X738, X729	
	Delivery of chemotherapy
Y136	Photodynamic therapy of lesion of organ

Table A.1.3 – OPCS codes that are not considered to represent recurrent H&N disease.

Procedure type	OPCS 4.4 codes	Description
Diagnostic and maintenance procedures of common non H&N malignant conditions	A841-A849	Neurophysiological operations
	L911-L916, L918, L919, L941-L949, L921-L923, L928, L929, L351-L353, L358, L359	Vein and artery-related operations
	M471-M475, M478, M479	Maintenance/operations of catheter

Table A.1.3 (*Continued*)

Procedure type	OPCS 4.4 codes	Description
	U181-U183, U188, U189	Imaging of breast
	U191-U199, U201-U206, U208, U209	Electrocardiography operations
	U301, U302, U308, U309	
	U311, U318, U319	Cardiovascular testing
	U321, U328, U329, X360-X363, X368, X369	Pacemaker testing
		Blood tests, blood withdrawal
Adjunctive procedures	F091-F095, F098, F099, F101-F104, F108, F109	Tooth extraction, dental clearance
	G342	Creation of temporary gastrostomy
		Blood transfusion, blood stem cell transplant
	X331-X336, X338, X339	
Reconstructive procedures	F081-F084, F088, F089	Implantation of tooth
	F631-F635, F638, F639	Fitting/insertion/adjustment of denture or obturator

Table A.1.4 – OPCS codes that represent a histological diagnosis procedure.

OPCS 4.4 codes	Description
C061, C106, C117, C222, C244, C263	Biopsy of eye region
D123, D201, D281	Biopsy of ear region
E033, E045, E095, E101, E134, E173, E202	Biopsy of nasal region/adenoids
E251, E252, E271, E334, E361, E368, E369, E371, E491	Biopsy of pharynx, nasopharynx and larynx, lower respiratory tract
F062, F203, F241, F321, F362, F421, F481	Biopsy of lip, gingiva, tongue, palate, tonsil, mouth, salivary gland
G131, G161, G191, G451	Biopsy of oesophagus, upper GI
J091, J092, J131, J132, J141, J171, J723	Biopsy of liver, spleen
S131, S132, S138, S139, S141-S144, S148, S149, S151, S152	Biopsy of skin/skin of head and neck region
T092, T141	Biopsy of pleura
T872	Biopsy of cervical lymph node
V194, V133	Biopsy of mandible, bone of face
E253, E258, E259	Endoscopic examination of nasopharynx

REFERENCES

[1] A. Jemal, F. Bray, J. Ferlay, Global cancer statistics, CA Cancer J. Clin. 61 (2) (2011) 69–90, http://dx.doi.org/10.3322/caac.20107.Available.

[2] National Head & Neck Cancer audit 2011. (2012). Retrieved from https://catalogue.ic.nhs.uk/publications/clinical/headneck/clin-audi-supp-prog-head-neck-dahn-2010-2011/clin-audi-supp-prog-head-neck-dahn-10-11-rep.pdf

[3] A.E. Powell, H.T.O. Davies, R.G. Thomson, Using routine comparative data to assess the quality of health care: understanding and avoiding common pitfalls, Qual. Saf. Health Care 12 (2) (2003) 122–128.

[4] Z.-W. Liu, H. Fitzke, M. Williams, Using routine data to estimate survival and recurrence in head and neck cancer: our preliminary experience in twenty patients, Clin. Otolaryngol. 38 (4) (2013) 334–339, http://dx.doi.org/10.1111/coa.12123.

[5] N.A. Christakis, E.B. Lamont, terminally ill patients: prospective cohort study, Br. Med. J. 320 (7233) (2000) 469–473.

[6] B.E. Kiely, Y.Y. Soon, M.H.N. Tattersall, M.R. Stockler, How long have I got? Estimating typical, best-case, and worst-case scenarios for patients starting first-line chemotherapy for metastatic breast cancer: a systematic review of recent randomized trials, J. Clin. Oncol. 29 (4) (2011) 456–463, http://dx.doi.org/10.1200/JCO.2010.30.2174.

[7] R Foundation for Statistical Computing, R: A Language and Environment for Statistical Computing, Vienna, Austria, 2013, Retrieved from http://www.r-project.org

[8] S.A.R. Nouraei, S. O'Hanlon, C.R. Butler, A. Hadovsky, E. Donald, E. Benjamin, G.S. Sandhu, A multidisciplinary audit of clinical coding accuracy in otolaryngology: financial, managerial and clinical governance considerations under payment-by-results, Clin. Otolaryngol. 34 (1) (2009) 43–51, http://dx.doi.org/10.1111/j.1749-4486.2008.01863.x.

[9] D.D. Dore, C. Liang, N. Ziyadeh, H. Norman, M. Bayliss, J.D. Seeger, Linkage of routinely collected oncology clinical data with health insurance claims data – an example with aromatase inhibitors, tamoxifen, and all-cause mortality, Pharmacoepidemiol. Drug Saf. 21 (2012) 29–36, http://dx.doi.org/10.1002/pds.3244.

[10] A. McColl, P. Roderick, C. Cooper, Hip fracture incidence and mortality in an English region: a study using routine National Health Service data, J. Public Health Med. 20 (2) (1998) 196–205.

[11] M.E. Stokes, D. Thompson, E.L. Montoya, M.C. Weinstein, E.P. Winer, C.C. Earle, Ten-year survival and cost following breast cancer recurrence: estimates from SEER-medicare data, Value Health 11 (2) (2008) 213–220, http://dx.doi.org/10.1111/j.1524-4733.2007.00226.x.

[12] B.L. Nordstrom, J.L. Whyte, M. Stolar, C. Mercaldi, J.D. Kallich, Identification of metastatic cancer in claims data, Pharmacoepidemiol. Drug Saf. 21 (S2) (2012) 21–28, http://dx.doi.org/10.1002/pds.3247.

[13] E.B. Lamont, J.E. Herndon, J.C. Weeks, I.C. Henderson, C.C. Earle, R.L. Schilsky, N.A. Christakis, Measuring disease-free survival and cancer relapse using Medicare claims from CALGB breast cancer trial participants (companion to 9344), J. Natl. Cancer Inst. 98 (18) (2006) 1335–1338, http://dx.doi.org/10.1093/jnci/djj363.