UNIVERSITY
OF WOLLONGONG
AUSTRALIA

University of Wollongong
Research Online

Illawarra Health and Medical Research Institute | Faculty of Science, Medicine and Health

2013

Ascertaining invasive breast cancer cases; the validity of administrative and self-reported data sources in Australia

Anna Kemp
University of Wollongong, akemp@uow.edu.au

David B. Preen
University of Western Australia

Christobel Saunders
University of Western Australia

C D'Arcy J. Holman
University of Western Australia

Max Bulsara
University of Notre Dame

See next page for additional authors

Publication Details

Kemp, A., Preen, D. B., Saunders, C., Holman, C. J., Bulsara, M., Rogers, K. & Roughead, E. E. (2013). Ascertaining invasive breast cancer cases; the validity of administrative and self-reported data sources in Australia. BMC Medical Research Methodology, 13 (1), 17-1-17-8.

Research Online is the open access institutional repository for the University of Wollongong. For further information contact the UOW Library: research-pubs@uow.edu.au

Ascertaining invasive breast cancer cases; the validity of administrative and self-reported data sources in Australia

Abstract
Background: Statutory State-based cancer registries are considered the 'gold standard' for researchers identifying cancer cases in Australia, but research using self-report or administrative health datasets (e.g. hospital records) may not have linkage to a Cancer Registry and need to identify cases. This study investigated the validity of administrative and self-reported data compared with records in a State-wide Cancer Registry in identifying invasive breast cancer cases. Methods: Cases of invasive breast cancer recorded on the New South Wales (NSW) Cancer Registry between July 2004 and December 2008 (the study period) were identified for women in the 45 and Up Study. Registry cases were separately compared with suspected cases ascertained from: i) administrative hospital separations records; ii) outpatient medical service claims; iii) prescription medicines claims; and iv) the 45 and Up Study baseline survey. Ascertainment flags included diagnosis codes, surgeries (e.g. lumpectomy), services (e.g. radiotherapy), and medicines used for breast cancer, as well as self-reported diagnosis. Positive predictive value (PPV), sensitivity and specificity were calculated for flags within individual datasets, and for combinations of flags across multiple datasets. Results: Of 143,010 women in the 45 and Up Study, 2039 (1.4%) had an invasive breast tumour recorded on the NSW Cancer Registry during the study period. All of the breast cancer flags examined had high specificity (>97.5%). Of the flags from individual datasets, hospital-derived 'lumpectomy and diagnosis of invasive breast cancer' and '(lumpectomy or mastectomy) and diagnosis of invasive breast cancer' had the greatest PPV (89% and 88%, respectively); the later having greater sensitivity (59% and 82%, respectively). The flag with the highest sensitivity and PPV ≥ 85% was 'diagnosis of invasive breast cancer' (both 86%). Self-reported breast cancer diagnosis had a PPV of 50% and sensitivity of 85%, and breast radiotherapy had a PPV of 73% and a sensitivity of 58% compared with Cancer Registry records. The combination of flags with the greatest PPV and sensitivity was '(lumpectomy or mastectomy) and (diagnosis of invasive breast cancer or breast radiotherapy)' (PPV and sensitivity 83%). Conclusions: In the absence of Cancer Registry data, administrative and self-reported data can be used to accurately identify cases of invasive breast cancer for sample identification, removing cases from a sample, or risk adjustment. Invasive breast cancer can be accurately identified using hospital-derived diagnosis alone or in combination with surgeries and breast radiotherapy.

Keywords
sources, australia, cases, cancer, validity, breast, self, invasive, administrative, ascertaining, reported, data

Disciplines
Medicine and Health Sciences

Publication Details
Kemp, A., Preen, D. B., Saunders, C., Holman, C. J., Bulsara, M., Rogers, K. & Roughead, E. E. (2013). Ascertaining invasive breast cancer cases; the validity of administrative and self-reported data sources in Australia. BMC Medical Research Methodology, 13 (1), 17-1-17-8.

This journal article is available at Research Online: http://ro.uow.edu.au/ihmri/337

Authors
Anna Kemp, David B. Preen, Christobel Saunders, C D'Arcy J. Holman, Max Bulsara, Kris Rogers, and Elizabeth E. Roughead

This journal article is available at Research Online: http://ro.uow.edu.au/ihmri/337

Kemp et al. BMC Medical Research Methodology 2013, 13:17
http://www.biomedcentral.com/1471-2288/13/17

BMC Medical Research Methodology

RESEARCH ARTICLE Open Access

Ascertaining invasive breast cancer cases; the validity of administrative and self-reported data sources in Australia

Anna Kemp[1,2*], David B Preen[1], Christobel Saunders[3], C D'Arcy J Holman[4], Max Bulsara[5], Kris Rogers[6] and Elizabeth E Roughead[7]

Abstract

Background: Statutory State-based cancer registries are considered the 'gold standard' for researchers identifying cancer cases in Australia, but research using self-report or administrative health datasets (e.g. hospital records) may not have linkage to a Cancer Registry and need to identify cases. This study investigated the validity of administrative and self-reported data compared with records in a State-wide Cancer Registry in identifying invasive breast cancer cases.

Methods: Cases of invasive breast cancer recorded on the New South Wales (NSW) Cancer Registry between July 2004 and December 2008 (the study period) were identified for women in the 45 and Up Study. Registry cases were separately compared with suspected cases ascertained from: i) administrative hospital separations records; ii) outpatient medical service claims; iii) prescription medicines claims; and iv) the 45 and Up Study baseline survey. Ascertainment flags included diagnosis codes, surgeries (e.g. lumpectomy), services (e.g. radiotherapy), and medicines used for breast cancer, as well as self-reported diagnosis. Positive predictive value (PPV), sensitivity and specificity were calculated for flags within individual datasets, and for combinations of flags across multiple datasets.

Results: Of 143,010 women in the 45 and Up Study, 2039 (1.4%) had an invasive breast tumour recorded on the NSW Cancer Registry during the study period. All of the breast cancer flags examined had high specificity (>97.5%). Of the flags from individual datasets, hospital-derived 'lumpectomy and diagnosis of invasive breast cancer' and '(lumpectomy or mastectomy) and diagnosis of invasive breast cancer' had the greatest PPV (89% and 88%, respectively); the later having greater sensitivity (59% and 82%, respectively). The flag with the highest sensitivity and PPV ≥ 85% was 'diagnosis of invasive breast cancer' (both 86%). Self-reported breast cancer diagnosis had a PPV of 50% and sensitivity of 85%, and breast radiotherapy had a PPV of 73% and a sensitivity of 58% compared with Cancer Registry records. The combination of flags with the greatest PPV and sensitivity was '(lumpectomy or mastectomy) and (diagnosis of invasive breast cancer or breast radiotherapy)' (PPV and sensitivity 83%).

Conclusions: In the absence of Cancer Registry data, administrative and self-reported data can be used to accurately identify cases of invasive breast cancer for sample identification, removing cases from a sample, or risk adjustment. Invasive breast cancer can be accurately identified using hospital-derived diagnosis alone or in combination with surgeries and breast radiotherapy.

Keywords: 45 and up study, Sensitivity, Specificity, Positive predictive value, Lumpectomy, Mastectomy, Radiotherapy, Hospital diagnosis, Tamoxifen, Anastrazole, Self-report

* Correspondence: anna.kemp@uwa.edu.au
[1]Centre for Health Services Research, School of Population Health, The University of Western Australia, 35 Stirling Hwy, Crawley, WA 6009, Australia
[2]Illawarra Health and Medical Research Institute, Building 32, University of Wollongong, Wollongong, NSW 2522, Australia
Full list of author information is available at the end of the article

BioMed Central

© 2013 Kemp et al.; licensee BioMed Central Ltd. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/2.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Background

Routinely-collected and self-reported health data are increasingly used to identify health status and service use in research. In Australia, State-based statutory cancer registries are considered the 'gold standard' for identifying breast cancer cases for research purposes and in recent years these data have been linked to other routinely-collected datasets for research [1-3].

Since December 2008, delays in release of mortality data from the Australian Bureau of Statistics have prevented the New South Wales (NSW) Cancer Registry from releasing data [4]. Consequently, the gold-standard dataset for identifying breast cancer in NSW has been inaccessible from 2009 onward and cancer researchers cannot ascertain cases from this source. Aside from these recent Australian issues, researchers in many countries face lengthy delays, cost or political barriers to accessing linked, routinely-collected datasets, which are often held by separate custodians and cover different jurisdictions [5-8]. Researchers who only have access to single datasets (e.g. hospital records), or specified packages of automatically linked datasets (e.g. English national hospital and death records [9], Australian medical service and prescription claims linked with NSW 45 and Up Study [10]) may want to identify cases of breast cancer without linkage to a Cancer Registry.

The aim of this study was to determine whether incident cases of invasive breast cancer can be accurately ascertained through a range of routinely-collected administrative and self-reported health datasets, with comparisons made to histologically-confirmed Cancer Registry records.

Methods

Study sample

The study sample was selected from participants enrolled in the 45 and Up Study; a cohort of approximately 267,000 adults aged ≥45 years residing in NSW [10]. Participants in this study provided demographic, lifestyle and health information upon joining the study and consented to having their routinely-collected health data linked and analysed for research purposes [11]. Baseline information for the 45 and Up Study cohort are already linked to medical service claims and pharmaceuticals publically-subsided by the Australian government. These datasets are now being used for many epidemiological studies e.g. [12-14]. Researchers can also apply to have these records linked to other NSW and national datasets on a project-by-project basis. Detailed information regarding the establishment and recruitment for the 45 and Up Study are described elsewhere [10]. The present study included 143,010 women recruited between January 2006 and April 2009, who had completed breast cancer-related items in the baseline survey of the 45 and Up Study.

Ascertaining cases using the gold standard

The NSW Cancer Registry contains, by statutory requirement, records of all cancers diagnosed or treated in NSW [15,16]. The Cancer Registry was considered the 'gold standard' source for cancer identification in this study. Cases were defined as women with a diagnosis of invasive breast cancer listed on the NSW Cancer Registry during the study period; 1 July 2004 to 31st December 2008. Codes used to identify cases were International Classification of Diseases version 10 with Australian modifications (ICD-10-AM) C50.0-C50.9 [17]. Participants with no registry record during the study period were considered non-cases.

Data sources and linkage

We accessed unit-record linked data from: i) the 45 and Up Study baseline survey, ii) NSW Admitted Patient Data Collection; iii) Medicare Benefits Schedule (MBS) claims; and iv) Pharmaceutical Benefits Scheme (PBS) claims. All data linkage was conducted by the NSW Centre for Health Record Linkage [18] and researchers were provided with de-identified data only.

Ascertaining cases using other datasets

Hospital diagnosis ascertainment flags were identified through the NSW Admitted Patient Data Collection. This dataset captures all admissions to public and private hospitals in the State of NSW. As with the Cancer Registry, we identified participants with a principal inpatient diagnosis of invasive breast cancer using ICD-10-AM codes C50.0-C50.9. Suspected cases flagged by inpatient diagnosis were defined as true positives if they occurred within three months of the Cancer Registry date of diagnosis. Flags for breast cancer surgeries were also identified from hospital data. We used ICD-10-AM procedure codes to identify mastectomy (31518–00, 31518–01, 31524–00, 31524–01), and excision of malignant breast lump (lumpectomy) (31500–00, 31500-01, 31503-00, 31503-01, 31506-00, 31506-01, 31509-00, 31509-01, 31512-00, 31512-01) [17]. Suspected cases flagged by surgeries occurring within three months of the Cancer Registry date of diagnosis were considered to be true positives.

Flags for breast radiotherapy and prescription medicines were identified through the MBS and PBS datasets. The MBS is a claims database which captures medical services subsidised by the Australian Federal Government for all Australian citizens [19]. As with the MBS, the PBS is a national scheme covering all Australian citizens [20]. Breast radiotherapy is conducted on an outpatient basis in NSW and was not detected in the hospital dataset. We identified claims for breast radiotherapy using MBS codes 15221, 15236, 15251, and 15266 [21]. We identified claims for dispensings of prescribed medicines used to treat breast cancer using PBS codes. These datasets captured the date of service for radiotherapy and dispensing of medicines.

These medicines included selective oestrogen reuptake inhibitors (tamoxifen 2109B, 2110C and toremifene 8216K), aromatase inhibitors (anastrozole 8179L, exemestane 8506Q, letrozole 8245Y); and other breast cancer therapies (goserelin 1452M; trastuzumab 4632T, 4639E, 4650R, 4703M, 7264H, 7265J, 7266K, 7267L; lapatinib 9148L; 500 mg preparations of medroxyprogesterone 2728N) [22]. All these therapies are only subsidized for use in women breast cancer. Only 500 mg preparations of medroxyprogesterone were included because lower dose preparations are subsidised for indications other than breast cancer in Australia [23]. Suspected cases of invasive breast cancer flagged by breast radiotherapy or the specified medicines were considered to be true positives if these services were provided within 12 months of the Cancer Registry date of diagnosis. The follow up periods for diagnosis, surgery, radiotherapy and prescription medicines vary and were selected to allow for the usual delays in treatment after diagnosis and were determined from sensitivity analysis examining different follow up periods (Additional file 1: Sensitivity analyses).

Process for comparing self-reported diagnosis with the cancer registry

Self-reported diagnosis of breast cancer was identified from the 45 and Up Study baseline survey. Recruitment to the study and completion of the baseline survey commenced in January 2006, making this the latest date where all participants would uniformly have the opportunity to self-report a diagnosis of breast cancer. Therefore, self-reports were only compared against cases in the Cancer Registry for the period 1 July 2004 to 31st December 2005.

The baseline survey asked participants to indicate their current age in years and months; whether a doctor had ever told them they had breast cancer (yes/no) and, if yes, their age in years at diagnosis. We then calculated the 12 month period in which the participant was the age they reported being at diagnosis. For example, a woman aged 72 years and 4 months when recruited to the study on 19th August 2008 and reporting a cancer diagnosis at age 68 would have a proxy 'diagnosis year' from April 2004 to March 2005. A true positive was defined as a self-reported diagnosis year overlapping the period July 2004-December 2005, and a Cancer Registry date of diagnosis during this period (see Figure 1, Participants A and B). A false positive was defined as occurring when the reported diagnosis year overlapped the period July 2004-December 2005 but no Cancer Registry record was found for the period (Participants C and D). A true negative was defined as occurring when the participant did not report a breast cancer diagnosis or reported a diagnosis year that did not overlap the period, and no Cancer Registry record was found for the period (Participants E and F). A false negative was defined as occurring when a participant did not report a diagnosis of breast cancer or reported a diagnosis year which did not overlap with the period and a Cancer Registry record was found for the period (Participants G and H).

Statistical analyses

Breast cancer flags drawn from individual datasets were compared against the Cancer Registry for PPV, sensitivity, and specificity (see Table 1) [24,25]. Researchers will prioritise these indicators differently depending on their

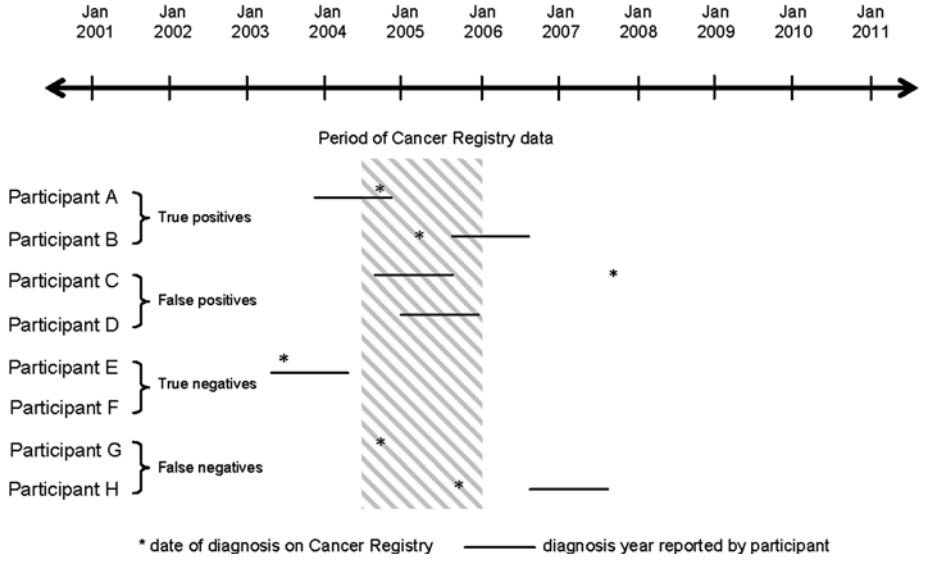

Figure 1 Possible relationships between the study period, diagnosis date recorded on the Cancer Registry, and the diagnosis year reported by participants.

Table 1 Validity of breast cancer flags in individual datasets compared with the Cancer Registry, July 2004-December 2008

Data set	Breast cancer flags	Positive predictive value	Sensitivity	Specificity
45 and Up baseline survey[1]	Self-reported diagnosis of breast cancer [2] within a year of birth-year reported	49.8%	84.7%	99.6%
Admitted Patient Data Collection	Diagnosis[3] of invasive breast cancer	85.9%	86.1%	99.8%
	Lumpectomy	52.0%	60.7%	99.2%
	Mastectomy	70.8%	32.6%	99.8%
	Lumpectomy OR mastectomy	56.4%	84.4%	99.1%
	Lumpectomy AND diagnosis of invasive breast cancer	89.0%	59.1%	99.9%
	Mastectomy AND diagnosis of invasive breast cancer	85.4%	31.8%	99.1%
	(Lumpectomy or mastectomy) AND diagnosis of invasive breast cancer	87.7%	82.3%	99.8%
	(Lumpectomy or mastectomy) OR diagnosis of invasive breast cancer	56.5%	88.2%	99.0%
	Mastectomy OR diagnosis of invasive breast cancer	79.7%	87.6%	99.7%
	Lumpectomy OR diagnosis of invasive breast cancer	58.2%	87.6%	99.1%
Medicare Benefits Schedule	Breast radiotherapy	72.8%	57.6%	99.7%
Pharmaceutical Benefits Scheme	Dispensed medicine for breast cancer [4]	45.5%	65.4%	98.9%

1: Comparison to Cancer Registry was restricted to July 2004-December 2005.
2: Reported cancer with birth-year occurring in or overlapping with the period July 2004-December 2005.
3: Primary diagnosis field.
4: Tamoxifen, toremifene, anastrozole, exemestane, letrozole, goserelin, trastuzumab, lapatinib, and medroxyprogesterone 500 mg.

reason for identifying cases (e.g. to define a sample, exclude cases from a sample, or for risk adjustment). We sought to identify flags with high PPV (≥85%) and within that, the greatest sensitivity. We intend to use these flags to identify a sample of women with invasive breast cancer for future studies.

To determine if combinations of flags improved PPV, sensitivity or specificity, clinically meaningful combinations of flags were determined in consultation with a breast cancer surgeon and medical oncologist. These combinations of flags (as described in Table 2) were derived from the commonly utilised combination of 45 and Up Study baseline, MBS and PBS data; and from all of the available datasets (hospital, 45 and Up Study baseline data, MBS and PBS data). These combination flags were also assessed against the Cancer Registry for PPV, sensitivity and specificity. Additional sensitivity analyses were undertaken to determine how many 'false reports' of breast cancer on the 45 and Up Study baseline survey were recorded cases on the Cancer Registry, but with incorrectly reported age at diagnosis. All analyses were conducted in IBM SPSS version 19.0.

Ethics approval

Ethical approval for this project was received from The University of Western Australia (WA) Human Research Ethics Committee (approval RA/4/1/4589) and the NSW Population and Health Services Research Ethics Committee (approval HREC/11/CIPHS/35).

Results

Of the 143,010 women in the 45 and Up Study cohort, 2039 (1.4%) had an invasive breast tumour recorded on the NSW Cancer Registry during the study period. Of these, 681 (33.4%) occurred between 1 July 2004 and 31st December 2005, and this subgroup was compared against self-reported breast cancer for the cohort.

Breast cancer flags from individual datasets

Table 1 shows the number of suspected cases flagged within each of the datasets examined. All of the breast cancer flags had high specificity (>98.5%). Self-reported diagnosis of breast cancer had a PPV of only 50% compared with the Cancer Registry; however the sensitivity was 85%. PPV and sensitivity of the hospital diagnosis of invasive breast cancer were both 86% and for lumpectomy 52% and 61%, respectively. Hospital-derived mastectomy had a higher PPV of 71% against the Cancer Registry but lower sensitivity (33%). When considering combinations of flags within hospital data, the one with the highest sensitivity for a flag with PPV over 85% was '(lumpectomy or mastectomy) and diagnosis of invasive breast cancer'. PPV and sensitivity for this combination were 88% and 82%, respectively. Among the flags from medical service and prescription medicine claims, breast

Table 2 Validity of combinations of breast cancer flags compared with the Cancer Registry, July 2004-December 2008

Breast cancer flags	Positive predictive value	Sensitivity	Specificity
45 and Up Study baseline, Medicare Benefits Schedule and Pharmaceutical Benefits Scheme			
Breast radiotherapy AND dispensed medicine	80.1%	40.3%	99.9%
Breast radiotherapy OR dispensed medicine[1]	47.9%	82.7%	98.7%
Breast radiotherapy AND self-reported diagnosis	69.3%	36.1%	99.9%
Breast radiotherapy AND dispensed medicine AND self-reported diagnosis[2]	72.1%	23.9%	99.9%
(Breast radiotherapy OR dispensed medicine) AND self-reported diagnosis	58.6%	60.4%	99.8%
Breast radiotherapy OR dispensed medicine OR self-reported diagnosis	27.0%	95.3%	97.7%
Admitted Patients Data Collection, 45 and Up Study baseline, Medicare Benefits Schedule and Pharmaceutical Benefits Scheme			
(Lumpectomy or mastectomy) AND diagnosis of invasive breast cancer AND breast radiotherapy	90.1%	47.5%	99.9%
(Lumpectomy or mastectomy) AND diagnosis of invasive breast cancer AND breast radiotherapy AND dispensed medicine	89.7%	33.7%	99.9%
(Lumpectomy or mastectomy) AND diagnosis of invasive breast cancer AND dispensed medicine	87.8%	55.0%	99.9%
(Lumpectomy or mastectomy) AND (diagnosis of invasive breast cancer OR breast radiotherapy)	83.0%	83.2%	99.8%
(Lumpectomy or mastectomy) AND (diagnosis of invasive breast cancer OR dispensed medicine)	84.1%	83.1%	99.8%
(Lumpectomy or mastectomy) AND (diagnosis of invasive breast cancer OR breast radiotherapy OR dispensed medicine)	80.5%	83.6%	99.7%
(Lumpectomy or mastectomy) AND diagnosis of invasive breast cancer AND self-reported diagnosis	88.5%	71.1%	99.9%
(Lumpectomy or mastectomy) AND diagnosis of invasive breast cancer AND breast radiotherapy AND self-reported diagnosis	87.3%	29.4%	99.9%

1: Tamoxifen, toremifene, anastrozole, exemestane, letrozole, goserelin, trastuzumab, lapatinib, and medroxyprogesterone 500 mg.
2: Flag combinations including self-reported diagnosis of breast cancer were compared against the Cancer Registry for the period July 2004 to December 2005.

radiotherapy had the highest PPV and sensitivity (73% and 58%, respectively). Use of any medicine for breast cancer had a PPV of 46% and sensitivity of 65%.

Breast cancer flags from multiple datasets

Combinations of flags derived from the package of 45 and Up Study baseline survey, MBS and PBS datasets are shown in Table 2. All of the ascertainment flags from multiple datasets had high specificity (>97.5%). None of the combinations of flags from these datasets had PPV >85%; the highest PPV being 80% for the flag combination 'breast radiotherapy and a dispensed medicine (sensitivity 40%). Very high sensitivity was observed for the flag combination of 'breast radiotherapy or a dispensed medicine or self-reported diagnosis of breast cancer' (95%); however PPV was low (27%).

Combinations of flags which included hospital data are also shown in Table 2. Specificity was above 99.5% for all the flag combinations. Good PPV (>85%) was found for several flag combinations including surgeries, hospital diagnosis, breast radiotherapy, dispensed medicines, or self-reported diagnosis; however sensitivity was lower (range 29%-71%). The combination of flags with the highest sensitivity and PPV over 85% was '(lumpectomy or mastectomy) and hospital diagnosis and self-reported diagnosis)' (PPV 89%, sensitivity 41%). In contrast to the flags derived from hospital data alone none of the combinations of flags from multiple datasets were found to have PPV ≥85% and sensitivity above 80%.

A total of 581 women were considered to have 'falsely' reported a diagnosis of breast cancer because they had no record of invasive breast cancer on the Cancer Registry within 12 months of the birth year they reported. Of these, 399 (69%) were found to have a Cancer Registry record for invasive breast cancer for an earlier or later period. These women had misreported their age at diagnosis but not their history of diagnosis.

Discussion

We sought to identify flags for invasive breast cancer with PPV ≥85% and, within that, the greatest sensitivity. Of the ascertainment flags examined from individual datasets, the flag meeting these criteria was hospital-derived 'diagnosis of invasive breast cancer'. When compared with the gold-standard Cancer Registry this flag combination had a PPV and sensitivity both of 86%. In other words, 86% of the suspected cases identified by this flag were true positives, and 86% of the cases listed on the Cancer Registry during the study period were identified by this flag. The addition of flags from other Australian datasets (i.e. medical service, prescription claims and survey data) to these hospital-

derived flags did not results in combinations with both PPV and sensitivity over 85%.

Researchers working with the combination of Australian medical service claims, pharmaceutical claims and self-reported data could most accurately identify cases of invasive breast cancer using the flag combination of 'breast radiotherapy and a dispensed medicine'. Around 80% of cases identified by this flag were true cases, compared with the gold standard, and this flag identified 40% of the invasive breast cancers recorded on the Cancer Registry during the study period. Much higher sensitivity was achieved with the flag 'breast radiotherapy or a dispensed medicine or self-reported diagnoses; however the corresponding PPV was poor (27%).

To our knowledge, this is the first study to examine the validity of multiple breast cancer flags from multiple datasets against an Australian State Cancer Registry. Such investigation is important due to the increasing use of administrative and self-reported data in epidemiological studies, and with the unavailability of Cancer Registry data in some jurisdictions. We have used health and medical records for a large, heterogeneous sample of women for whom all public and private inpatient diagnoses and surgeries, subsidised outpatient procedures and medicines have been captured.

Some limitations exist which may have implications for this study. This study was conducted as part of a larger program of research examining use of endocrine therapies for invasive breast cancer in Australian clinical practice. The data we requested from the Cancer Registry were therefore restricted to invasive breast cancer and did not include records for ductal carcinoma in situ (DCIS). We were therefore unable to determine how often false positive flags were picking up genuine cases of DCIS and how many were unrelated to breast cancer of any kind. We examined the validity of various breast cancer flags for women in the 45 and Up Study who, by definition, are aged 45 years and over and have consented to their health records being used for research purposes. The health service use of these women may differ from younger women with breast cancer, or women who do not agree to participate in cohort studies. Therefore, the PPV, sensitivity and specificity calculated here for various flags may differ from those that would be found in whole-of-population studies. The validity of the flags examined here are impacted by the proportion of women who move out of NSW between diagnosis and treatment, as well as those dying prior to treatment or declining treatment. It may also be that the validity of the breast cancer flags examined here will change over time in response to changes in health service use and medical advancement.

Each of the flags we examined had very high specificity, which is to be expected given the low prevalence of breast cancer within the cohort (1.4%). In such a scenario, even a model which predicted no breast cancer at all would retain high specificity. Therefore, it is important to examine the PPV and sensitivity of all predictors. The optimum method for identifying cases of breast cancer without access to a Cancer Registry will depend on the type and number of datasets available and the reason cases need to be identified. Researchers seeking to exclude possible cases of breast cancer from their datasets will be most concerned with the specificity of breast cancer flags. All of the breast cancer flags we examined in this study, whether derived from individual or multiple datasets, had high specificity (>97.5%). Each of these would be suitable for identifying non-cases with high accuracy. Researchers wishing to identify any suspected cases of breast cancer for situations where some false positives are acceptable, such as risk adjustment, would likely prioritise flags with high sensitivity. In contrast, PPV would likely be most important for researchers seeking to identify breast cancer cases with the fewest possible false negatives (e.g. to select an affected cohort) [26].

The sensitivity and specificity of the hospital-derived flags we calculated are similar to those reported in a NSW study, which demonstrated the hospital procedures 'lumpectomy or mastectomy' identified invasive breast cancers in the Cancer Registry with high sensitivity (83%) and specificity (95%) [27]. International studies have also reported high accuracy for hospital records in identifying breast cancer [26,28,29]. In an Italian study of hospital records, the combination of hospital diagnosis together with 'lumpectomy or mastectomy' accurately identified the majority of cases on the Cancer Registry (PPV 91%, sensitivity 85%, specificity 99%) [26].

We found that self-reported diagnosis of breast cancer correctly identified 50% of invasive breast cancer diagnoses to within 12-months of the birth year reported. While one would expect individuals to self-report diagnoses such as cancer reliably [30,31], the baseline survey did not ask woman to differentiate between invasive breast cancer and DCIS. Women may have accurately reported a DCIS as a diagnosis of breast cancer, however our data extract from the Cancer Registry was limited to invasive tumours so this was not able to be confirmed. In addition, women may not accurately recall the age at which they were diagnosed [30-32]. In this study, women reporting a 'diagnosis year' overlapping the period July 2004 to December 2005 but without a Cancer Registry diagnosis during this period were considered false positives. A sensitivity analysis indicated that 399 of 581 (69%) of these 'false positives' (according to our definition) did have a Cancer Registry diagnosis for invasive breast cancer, but had incorrectly reported their age at diagnosis.

Conclusion

The Cancer Registry is the gold standard for identifying incident cases of invasive breast cancer in most jurisdictions.

The findings from this study indicate that other administrative and self-reported datasets examined can be used to accurately identify cases of invasive breast cancer when Cancer Registry data are unavailable. Cases of invasive breast cancer were most accurately identified by hospital-derived diagnosis of invasive breast cancer. This flag would be most suitable for researchers seeking to identify a study cohort with invasive breast cancer or for risk adjustment [26]. However, all of the flags examined in this study accurately identified cases without invasive breast cancer, so are suitable for researchers wishing to exclude cases from population-based datasets likely to have low prevalence of breast cancer.

Additional file

Additional file 1: Sensitivity analyses comparing follow-up periods for selected flags, compared with the Cancer Registry, July 2004-December 2008.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

AK conceived of the study, performed the statistical analyses, interpreted data and drafted the manuscript. ER participated in the design of the study and interpretation of data and helped to write the manuscript. KR and MB aided with data interpretation and critically reviewed the manuscript. DP, CS and CDH assisted with acquisition of data and critically reviewed the manuscript. All authors read and approved the final manuscript.

Acknowledgments

The authors are grateful to the 45 and Up Study participants. The 45 and Up Study is managed by The Sax Institute in collaboration with major partner Cancer Council New South Wales; and partners the National Heart Foundation of Australia (NSW Division); NSW Ministry of Health; *beyondblue: the national depression initiative*; Ageing, Disability and Home Care, NSW Family and Community Services; the Australian Red Cross Blood Service; and UnitingCare Ageing. The authors wish to thank staff at the NSW Centre for Health Record Linkage, as well as the data custodians NSW Cancer Institute, NSW Ministry of Health, and the Commonwealth Department of Human Services. This study was jointly-funded by Cancer Australia and the National Breast Cancer Foundation; which included salary support for AK and MB. LR is supported by an Australian Research Council Future Fellowship. The funding bodies did not contribute to the design, collection, analysis, or interpretation of data; manuscript writing; or the decision to submit the manuscript for publication.

Author details

[1]Centre for Health Services Research, School of Population Health, The University of Western Australia, 35 Stirling Hwy, Crawley, WA 6009, Australia. [2]Illawarra Health and Medical Research Institute, Building 32, University of Wollongong, Wollongong, NSW 2522, Australia. [3]School of Surgery, The University of Western Australia, 35 Stirling Hwy, Crawley, WA 6009, Australia. [4]School of Population Health, The University of Western Australia, 35 Stirling Hwy, Crawley, WA 6009, Australia. [5]Institute of Health and Rehabilitation Research, University of Notre Dame, PO Box 1225, Fremantle, WA 6959, Australia. [6]Prevention Research Collaboration, Sydney School of Public Health, University Fisher Road, Sydney, NSW 2206, Australia. [7]Quality Use of Medicines and Pharmacy Research Centre, School of Pharmacy and Medical Sciences, University of South Australia, GPO Box 2471, Adelaide, SA 5001, Australia.

Received: 17 October 2012 Accepted: 1 February 2013
Published: 11 February 2013

References

1. Fritschi L, Dye SA, Katris P: **Validity of melanoma diagnosis in a community-based screening program.** *Am J Epidemiol* 2006, **164:**385–390.
2. Stavrou E, Vajdic C, Loxton D, Pearson SA: **The validity of self-reported cancer diagnoses and factors associated with accurate reporting in a cohort of older Australian women.** *Cancer Epidemiol* 2011, **35:**75–80.
3. Stavrou EP, Lu CY, Buckley N, Pearson S: **The role of comorbidities on the uptake of systemic treatment and 3-year survival in older cancer patients.** *Ann Oncol* 2012, **23:**2422–2428.
4. *Accessing our data*. http://www.cancerinstitute.org.au/data-and-statistics/accessing-our-data.
5. Cui X, Werk C, Twilley L: *Learnings from anonymous data linkage acorss government sectors*. Perth, Australia: Presented at the International Data Linkage Conference 2012; 2012.
6. Stone CA, Ramsden CA, Howard JA, Roberts M, Halliday JH: *From data linkage to policy and back again: A preliminary report on the linkage of neonatal data in Victoria*. Sydney, Australia: Symposium on Health Data Linkage; 2000:209–211.
7. Goddard M, Mannion R, Smith P: **Enhancing performance in health care: a theoretical perspective on agency and the role of information.** *Health Econ* 2000, **9:**95–107.
8. Bradley CJ, Penberthy L, Devers KJ, Holden DJ: **Health services research and data linkages: issues, methods, and directions for the future.** *Health Serv Res* 2010, **45:**1468–1488.
9. Gill L, Goldacre M: *English national record linkage of hospital episode statistics and death registration records*. Oxford: Unit of Health-Care Epidemiology, Oxford University; 2003.
10. 45 and Up Study Collaborators: **Cohort profile: the 45 and up study.** *Int J Epidemiol* 2008, **37:**941–947.
11. *Information for researchers: the 45 and Up Study*. http://www.45andup.org.au/applyingtousethestudyresource.aspx.
12. Rogers K, Kemp A, McLachlan A, Blyth F: *Patterns of prescription opioid use for non-cancer pain in Australia: findings from the New South Wales 45 and Up Study*. Perth, Australia: International Data Linkage Conference 2012; 2012.
13. Goldsbury D, Harris MF, Pascoe S, Olver I, Barton M, Spigelman A, O'Connell D: **Socio-demographic and other patient characteristics associated with time between colonoscopy and surgery, and choice of treatment centre for colorectal cancer: a retrospective cohort study.** *BMJ Open* 2012, doi:10.1136/bmjopen-2012-001070.
14. Douglas K, Yen L, Korda R, Kljakovic M, Glasgow N: **Chronic disease management items in general practice: a population-based study of variation in claims by claimant characteristics.** *Med J Aust* 2011, **195:**198–202.
15. Cancer Institute of New South Wales: Sydney: Cancer Institute of NSW; http://www.cancerinstitute.org.au/cancer_inst/publications/journal_articles.html#2009.
16. *NSW Central Cancer Registry*. http://www.cancerinstitute.org.au/cancer_inst/statistics/registry.html.
17. National Centre for Classification in Health: *The international statistical classification of diseases and related health problems, 10th revision, Australian modification (ICD-10-AM)*. Sydney: National Centre for Classification in Health; 2006.
18. *Information for researchers*. http://www.cherel.org.au.
19. *Medicare*. http://www.humanservices.gov.au/customer/services/medicare/medicare.
20. *About the PBS*. http://www.pbs.gov.au/info/about-the-pbs.
21. *Schedule of Medicare Benefits, 1 July 2012*. http://www.health.gov.au/internet/mbsonline/publishing.nsf/Content/Medicare-Benefits-Schedule-MBS-1.
22. *PBS for health professionals*. http://www.pbs.gov.au/html/healthpro/home.
23. *Schedule of Pharmaceutical Benefits, 1 July 2012*. http://www.pbs.gov.au/pbs/home.
24. Altman DG, Bland JM: **Statistics notes: diagnostic tests 1: sensitivity and specificity.** *BMJ* 1994, **308:**1552.
25. Altman DG, Bland JM: **Statistics notes: diagnostic tests 2: predictive values.** *BMJ* 1994, **309:**102.
26. Yuen E, Louis D, Cisbani L, Rabinowitz C, De Palma R, Maio V, Leoni M, Grilli R: **Using administrative data to identify and stage breast cancer cases: implications for assessing quality of care.** *Tumori* 2011, **97:**428–435.
27. McGeechan K, Kricker A, Armstrong B, Stubbs J: **Evaluation of linked cancer registry and hospital records of breast cancer.** *Aust N Z J Public Health* 1998, **22:**765–770.
28. Nattinger AB, Laud PW, Bajorunaite R, Sparapani RA, Freeman JL: **An algorithm for the use of medicare claims data to identify women with incident breast cancer.** *Health Serv Res* 2004, **39:**1733–1749.

29. Freeman JL, Zhang D, Freeman DH Jr, Goodwin JS: **An approach to identifying incident breast cancer cases using medicare claims data.** *J Clin Epidemiol* 2000, **53:**605–614.
30. Bergmann MM, Byers T, Freedman DS, Mokdad A: **Validity of self-reported diagnoses leading to hospitalization: a comparison of self-reports with hospital records in a prospective study of american adults.** *Am J Epidemiol* 1998, **147:**969–977.
31. Desai M, Bruce ML, Desai R, Druss B: **Validity of self-reported cancer history: a comparison of health interview data and cancer registry records.** *Am J Epidemiol* 2001, **153:**299–306.
32. Abraham L, Geller BM, Yankaskasc B, Bowlesa E, Karlinerd L, Tapline T, Migliorettia D: **Accuracy of self-reported breast cancer among women undergoing mammography.** *Breast Cancer Res Treat* 2009, **118:**583–592.

doi:10.1186/1471-2288-13-17
Cite this article as: Kemp *et al.*: Ascertaining invasive breast cancer cases; the validity of administrative and self-reported data sources in Australia. *BMC Medical Research Methodology* 2013 **13**:17.

Submit your next manuscript to BioMed Central and take full advantage of:

- **Convenient online submission**
- **Thorough peer review**
- **No space constraints or color figure charges**
- **Immediate publication on acceptance**
- **Inclusion in PubMed, CAS, Scopus and Google Scholar**
- **Research which is freely available for redistribution**

Submit your manuscript at
www.biomedcentral.com/submit

BioMed Central