Short report

# Risks of using medical record and administrative data for prognostic models

Marliese Alexander[1,2], Sue M Evans[2], Rory Wolfe[2], David L Ball[1,3], Kate Burbury[1]

Retrospective epidemiological data derived from medical record (MR) or administrative data (AD) can be used to inform prognostic modelling. However, the validity of such projections depends on the data source accurately collecting details on patients' health status. We extracted data from three independent data sources — MR (patient records), AD (International Classification of Diseases [ICD] discharge coding), and prospective pre-defined cohort study data (SD) — for a randomly selected sample of 111 patients (10%) from the 1090 Australians diagnosed with lung cancer during 2012–2015 included in the Thoracic Malignancies Cohort (TMC). This information was used for prognosis estimation, including generating the simplified comorbidity score (SCS).[1] Ethics approval was provided by the Peter MacCallum Cancer Centre (reference, 14/107) and Monash University ethics committees (reference, CF15/727–2015000333).

The reporting of prognostic factors, including SCS comorbidities, was assessed in all three sources for completeness and inter-source agreement ($\kappa$ statistic). The baseline characteristics of our sample group were similar to those of the entire TMC. AD did not provide relevant prognostic data for 47 patients (42%) treated in ambulatory settings, and for 64 it was inaccurate. For example, 15 of 64 patients (23%) had at least one comorbidity according to AD, compared with 75 (68%) according to MR and 71 (64%) according to SD data; 24 patients (38%) had a positive smoking history, compared with 87 (78%) in MR and 90 (81%) in SD data; one patient (2%) had a respiratory comorbidity, compared with 31 (28%) in MR and 41 (37%) in SD data. Similar patterns applied to other SCS contributors. The perhaps most important prognostic factor, TNM staging, was recorded for 50 patients (45%) in MR data at the time of first treatment, and the concordance with SD data was good ($\kappa = 0.9$; 95% confidence interval [CI], 0.7–1.0). The most frequently documented factors in MR data included smoking status (completeness, 96%; $\kappa = 0.9$, 95% CI, 0.8–1.0), performance status (completeness, 82%; $\kappa = 0.5$, 95% CI, 0.4–0.7), and weight loss (completeness, 71%; $\kappa = 0.3$, 95% CI, 0.1–0.5). The median SCS, derived from SD and MR data, was 8 ($\kappa = 0.5$ [ie, moderate agreement], 95% CI, 0.4–0.7), and 1 when derived from AD (compared with SD: $\kappa = 0.3$ [poor agreement], 95% CI, 0.2–0.4) (Box).

Poor capture of the factors required for accurately estimating lung cancer prognosis limits the value of AD for clinical research. Deficits have also been reported for other lung cancer cohorts, with MR data as the primary comparator.[2,3] However, we also found deficiencies and inconsistencies in MR data. Without being able to accurately adjust for risk factors, health outcomes cannot be meaningfully compared. Divergent outcomes of prognostic models, including the SCS,[4] may be partially explained by differences in data acquisition methods.

Unreliable data capture reflects the varying purposes of these datasets. The MR is a real time communication tool for directing care pathways, therapy, and subsequent monitoring, and is often limited to immediately relevant clinical elements. AD capture details on health care use and funding. Coding of comorbidities reflects resource needs during hospitalisation;[5] their contribution to disease progression, morbidity, and competing mortality risk is less clearly projected. AD is particularly limited for ambulatory care models, of increasing importance for patients with either malignant or non-malignant diseases.

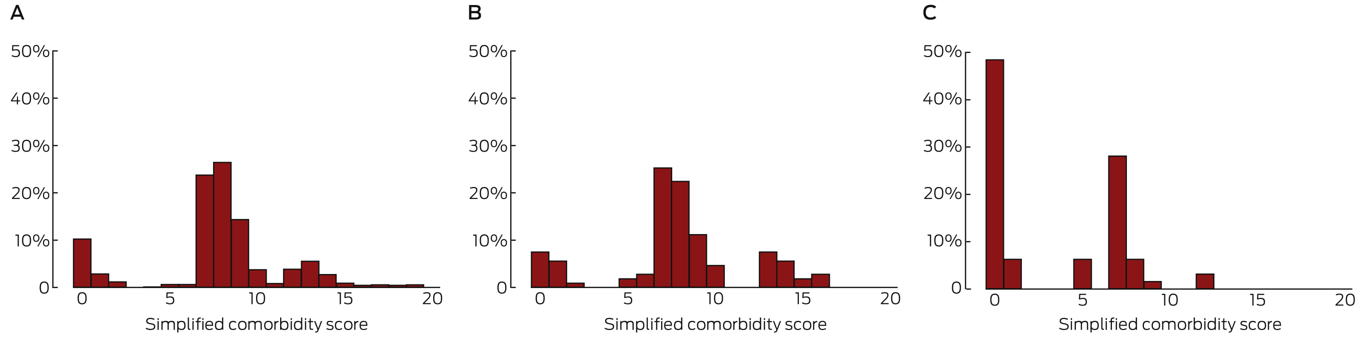
The capture of key information by AD and MRs for the diagnosis, prognosis and classification of lung cancer was suboptimal, and neither had the necessary detail for prognostication or longer term monitoring of health outcomes.

**Acknowledgements:** Marliese Alexander is a doctoral (PhD) candidate and was supported by an Australian Postgraduate Award from the Department of Epidemiology and Preventive Medicine, Monash University. Sue Evans is supported by a Monash Partners Academic Health Science Centre Clinical Research Fellowship.

**Competing interests:** No relevant disclosures. ■

© 2017 AMPCo Pty Ltd. Produced with Elsevier B.V. All rights reserved.

Simplified comorbidity score (SCS) distribution, as derived from (A) prospective study data, (B) medical records, and (C) administrative data*



* Includes only the 64 patients for whom data were available in all three sources, to ensure unbiased comparison of SCS. ◆

1 Colinet B, Jacot W, Bertrand D, et al. A new simplified comorbidity score as a prognostic factor in non-small-cell lung cancer patients: description and comparison with the Charlson's index. *Br J Cancer* 2005; 93: 1098-1105.

2 Kehl KL, Lamont EB, McNeil BJ, et al. Comparing a medical records-based and a claims-based index for measuring comorbidity in patients with lung or colon cancer. *J Geriatr Oncol* 2015; 6: 202-210.

3 Seo HJ, Yoon SJ, Lee SI, et al. A comparison of the Charlson comorbidity index derived from medical records and claims data from patients undergoing lung cancer surgery in Korea: a population-based investigation. *BMC Health Serv Res* 2010; 10: 236-243.

4 Alexander M, Evans SM, Stirling RG, et al. The influence of comorbidity and the simplified comorbidity score on overall survival in non-small cell lung cancer-a prospective cohort study. *J Thorac Oncol* 2016; 11: 748-757.

5 National Centre for Classification in Health. Australian coding standards for ICD-10-AM and the Australian Classification of Health Interventions. Seventh edition. Sydney: NCCH, University of Sydney, 2010. ■

[1] Peter MacCallum Cancer Centre, Melbourne, VIC. [2] Monash University, Melbourne, VIC. [3] University of Melbourne, Melbourne, VIC. Marliese.Alexander@petermac.org

▪ doi: 10.5694/mja16.00919