Quality in Primary Care 2008;16:49–52

Short report

Validation of THIN data for non-melanoma skin cancer

Andy Meal BMedSci (Hons) BM BS MPhil FHEA
Lecturer, School of Nursing

Jo Leonardi-Bee BSC (Hons) MSc PhD
Lecturer in Medical Statistics, Division of Epidemiology and Public Health

Chris Smith BA PGCE
Senior Research Fellow, Division of Epidemiology and Public Health

Richard Hubbard MSc DM FRCP
British Lung Foundation Professor of Respiratory Epidemiology, Division of Epidemiology and Public Health

Fiona Bath-Hextall BSc (Hons) PhD
Associate Professor, Centre for Evidence-Based Dermatology

University of Nottingham, UK

ABSTRACT

Background The Health Improvement Network (THIN) database began in 2003. It consists of anonymised records from over 300 general practice computer systems and is likely to be valuable for research, planning and strategic issues in health care, but it is important to establish completeness and accuracy of the data.

Aim To investigate the validity of THIN data for non-melanoma skin cancer (NMSC). We defined NMSC as basal cell carcinoma (BCC) and squamous cell carcinoma (SCC).

Methods Using Read codes we extracted THIN database records of first-recorded diagnoses of NMSC from 1 January 1996 to 31 December 2003. Searches for SCC were unable to distinguish between skin tumours of this type, and SCC at any other site. From our dataset for BCC, 40 patient records were selected at random, and a questionnaire sent to their corresponding practice, asking if they had been referred to hospital/dermatology clinic, and how the diagnosis of BCC had been confirmed.

Results All the patients in the sample were referred to a hospital or dermatology clinic: 37/40 (93%) had the diagnosis of BCC confirmed, either by a letter from the hospital or a pathology report, a finding that we have reported previously. One patient's diagnosis was confirmed as SCC, and the other two either died or moved away before diagnosis could be confirmed. The 38 patients with diagnoses confirmed were all treated in hospital or dermatology clinic.

Conclusions Data for BCC are sufficiently accurate for research. It is also likely that these data will prove valuable for quality management. It is not possible currently to obtain accurate data for SCC of the skin from the THIN database. This seems not to be a problem with the THIN database itself, but attributable to the Read coding scheme being, in practice, unable to allow differentiation between SCCs of different organs.

Keywords: database, non-melanoma skin cancer, THIN, validation

How this fits in with quality in primary care

What do we know?
THIN is a new aggregated database of anonymised patient records from over 300 general practices. Non-melanoma skin cancer is an increasingly common presentation in primary care.

What does this paper add?
THIN data for basal cell carcinoma are sufficiently accurate for research. Accurate data for squamous cell carcinoma of the skin cannot be obtained due to the structure of the Read code system.

Introduction

General practice computer systems offer access to a valuable data source.[1] Research, planning and strategic issues in healthcare delivery could find these data particularly valuable.[2] Aggregated databases created from electronic patient records in general practice have become an important source of data for research, but completeness and accuracy of these data are essential.[3] Recently The Health Improvement network (THIN), a collaboration between In Practice Systems Ltd (InPS) and EPIC Database and Research Company Ltd, has created a research database of anonymised patient records from information entered by general practices in their ViSion computer systems.[4] Data collection began in 2003, and currently the database contains data from over 300 practices, most of these having over 15 years of data on their computer systems.[5] The rationale and conceptual background to this database are described by Bourke *et al*.[6] Lewis *et al* provide more background information in their recent validation study.[3] In particular, they point out that there is some overlap between the THIN data and the well-established General Practice Research Database (GPRD) data in that some practices contribute to both schemes, but that some data in the THIN database are not derived from existing GPRD practices. They conclude, however, that THIN data collected outside GPRD appear to be as valid as the data collected as part of GPRD.

We studied the descriptive epidemiology, particularly incidence trends, in non-melanoma skin cancer (NMSC). We included only squamous cell carcinoma (SCC) and basal cell carcinoma (BCC) in our definition of NMSC. We report here on the validity of our data, in particular whether lesions coded as BCC are confirmed as BCC in secondary care, and the implications this might have for future studies of NMSC. A paper that illustrates our epidemiological data for BCC has been published elsewhere.[7]

Methods

Data were extracted from the THIN database records for first recorded diagnoses of NMSC from 1 January 1996 to 31 December 2003. At the time of our data extraction, five-byte Read codes were used to record data on the THIN database. Our Read code searches for SCC were unable to distinguish between skin tumours of this type and SCC at any other site. We therefore chose not to progress further with our analyses of records of SCC and instead concentrated on BCC. The Read codes used were B33..11 (basal cell carcinoma), B33..13 (rodent ulcer), and B33..16 (epithelioma basal cell). From our dataset,[7] 40 patient records were selected using a computer-generated pseudo-random list using Stata, and a questionnaire sent to the practice where each patient was registered. We assessed validity of our data by asking if the patient had been referred to a hospital/dermatology clinic, and how the diagnosis of BCC had been confirmed. Once our questionnaire had been created, it was distributed to the resulting 22 practices by EPIC. We, as researchers, did not contact the practices directly, nor were we aware of the identity of the practices. Practices returned the completed questionnaires to EPIC, who in turn returned them to the lead investigator (FB-H).

Data are presented as means and standard deviations (SD) for continuous data, and as numbers and percentages (%) for categorical data.

Results

Out of our total of 40 patients, 21 were male and 19 were female. Mean age at event of the males was 66.9 years (S.D. 11.3 years), and for the females was 78.6 years (SD 7.9 years). Twenty-two of the 40 patients had one recording of BCC and the rest had multiple recordings within the time-frame of the

study. All the patients were referred to a hospital or dermatology clinic. This confirmed that our sample of patients with BCC had a corresponding record in the relevant general practice. Thirty-seven out of forty (93%) had the diagnosis of BCC confirmed, either by a letter from the hospital or a pathology report, a finding that we have reported previously.[7] Of the three patients whose diagnosis of BCC was not confirmed, one was confirmed as SCC, and the other two either died or moved away before diagnosis could be confirmed. This led us to believe that our sample was made up of accurate records of BCC. The 38 patients with diagnoses confirmed, including the patient with SCC, were all treated in hospital or dermatology clinic. For the 37 patients with confirmed BCC, 34 (92%) received surgical treatment and the remaining three received non-surgical treatment (see Box 1).

Box 1 Summary of questionnaire findings for sample of records of BCC

Sample size	40
Referred to hospital	40
Treated in hospital	38
Confirmed as BCC	37
Confirmed as SCC	1
Lost to follow-up	2

Discussion

To our knowledge, there has been no other study of NMSC epidemiology using THIN data. Therefore our experiences of using this database should be relevant to others wishing to undertake research in this area, or to healthcare providers who want to investigate service provision in this important and increasingly common form of cancer. Our main findings are that the data for BCC are sufficiently accurate for research purposes. The high percentage of patients with hospital-confirmed diagnoses and treatment supports earlier findings that only between 1.3% and 8.8% of BCCs are managed in primary care.[8] Therefore our findings should also be relevant to service provision in hospital care for NMSC. By extension, we feel that it is likely that these data will prove valuable for quality management. In contrast to this, our experience suggests that it is not possible currently to obtain accurate data for SCC of the skin from the THIN database. This seems not to be a problem with the THIN database itself, but attributable to the Read coding scheme being, in practice, unable to allow differentiation between SCCs of different organs. Not only does this make research into this subject difficult, but it is likely also to be an issue for audit and other forms of quality management, since our experiences suggest that the hospital-confirmed diagnoses of NMSC are classified by morphology.

This is a small validation study of one cancer, therefore we cannot infer validity of THIN data for any other diagnoses. In common with other studies that use data from GP databases, we cannot know if the total number of BCCs is a true reflection of the number in the general population. Only people with a BCC who consult will be recorded in the database. Furthermore, we have no way of knowing if any BCCs are recorded as a different diagnosis in the database. Inspite of these limitations we believe that our study shows it is likely that BCCs recorded in the THIN database are accurate, confirmed diagnoses.

ACKNOWLEDGEMENTS

We would like to thank the staff at EPIC and InPS for help and advice in accessing the THIN database and distributing the validation questionnaires.

FUNDING

The University of Nottingham School of Nursing, and Institute of Clinical Research.

ETHICS

Our study had approval from EPIC (a research organisation facilitating the research use of electronic databases of primary care records from the UK National Health Service), Nottingham Ethics Committee (04/Q2403/140) and Queen's Medical Centre Medical Research and Development Committee, Nottingham.

REFERENCES

1 Pringle M and Hobbs R. Large computer databases in general practice. *BMJ* 1991;302:741–2.
2 Hammersley V, Meal A, Wright L and Pringle M. Using MIQUEST in general practice. *Journal of Informatics in Primary Care* 1998;November:3–7.
3 Lewis J, Schinnar R, Biker W, Wang X and Strom B. Validation studies of the Health Improvement Network (THIN) database for pharmacoepidemiology research. *Pharmacoepidemiology and Drug Safety* 2007;16:393–401.
4 The Health Improvement Network (THIN) www.thin-uk.com (accessed 21 November 2007).
5 EPIC. *THIN – Data Collection.* www.epic-uk.org/thin_data_collection.htm (accessed 21 November 2007).
6 Bourke A, Dattani H and Robinson M. Feasibility study and methodology to create a quality-evaluated database of primary care data. *Informatics in Primary Care* 2004;12:171–7.
7 Bath-Hextall F, Leonardi-Bee J, Smith C, Meal A and Hubbard R. Trends in incidence of skin basal cell carcinoma. Additional evidence from a UK primary care database study. *International Journal of Cancer* 2007;121(9):2105–8.

8 National Institute for Health and Clinical Excellence. *Guidance on Cancer Services. Improving Outcomes for People with Skin Tumours Including Melanoma: evidence review*. London: NICE, 2006.

CONFLICTS OF INTEREST

None.

ADDRESS FOR CORRESPONDENCE

Dr Andy Meal, School of Nursing, Medical School, University of Nottingham, Nottingham NG7 2HA, UK. Tel: +44 (0)115 8230903; fax +44 (0)115 8230999; email: Andy.Meal@nottingham.ac.uk

Received 1 August 2007
Accepted 24 September 2007