6. Elley CR, Kenealy T, Robinson E, Drury PL. Glycated haemoglobin and cardiovascular outcomes in people with Type 2 diabetes: a large prospective cohort study. *Diabet Med* 2008; **25(11):** 1295–1301.
7. Khaw KT, Wareham N, Luben R, *et al.* Glycated haemoglobin, diabetes, and mortality in men in Norfolk cohort of european prospective investigation of cancer and nutrition (EPIC-Norfolk). *BMJ* 2001; **322(7277):** 15–18.
8. Action to Control Cardiovascular Risk in Diabetes Study Group. Effects of intensive glucose lowering in Type 2 diabetes. *N Engl J Med* 2008; **358(24):** 2545–2559.
9. ADVANCE Collaborative Group. Intensive blood glucose control and vascular outcomes in patients with type 2 diabetes. *N Engl J Med* 2008; **358(24):** 2560–2572.
10. Duckworth W, Abraira C, Moritz T, *et al.* Glucose control and vascular complications in veterans with type 2 diabetes. *N Engl J Med* 2009; **360(2):** 129–139.
11. Holman RR, Paul SK, Bethel MA, *et al.* 10-year follow-up of intensive glucose control in type 2 diabetes. *N Engl J Med* 2008; **359(15):** 1577-1589.
12. Currie CJ, Peters JR, Tynan A, *et al.* Survival as a function of HbA_{1c} in people with type 2 diabetes: a retrospective cohort study. *The Lancet* 2010; **375(9713):** 481–489.
13. American Diabetes Association. Standards of Medical Care in Diabetes — 2010. *Diabetes Care* 2010; **33(Suppl 1):** S11–S61.
14. National Institute for Health and Clinical Excellence. Type 2 diabetes: newer agents for blood glucose control: costing report. London: NICE, 2009. http://guidance.nice.org.uk/CG87/CostReport/pdf/English (accessed 10 Feb 2010).

DOI: 10.3399/bjgp10X483463

ADDRESS FOR CORRESPONDENCE

Jonathan Graffy
General Practice and Primary Care Research Unit, University of Cambridge, Institute of Public Health, Forvie Site, Robinson Way, Cambridge, CB2 0SR, UK.
E-mail: jonathan.graffy@phpc.cam.ac.uk

Case validation in research using large databases

Computerised healthcare databases (CHCDs) have been increasingly used in epidemiologic research and have become the single most used source of information in pharmacoepidemiology. A key feature in the selection of a computerised database for research is completeness and validity of the data. As Khan *et al* highlight in the present issue of the *BJGP*,[1] researchers should investigate their information source and how well it covers the diagnosis under study.

The validation process of a database is complex, and the resources required to implement a study protocol using CHCDs will vary widely depending on the need and amount of validation required. It all starts with the selection of a database with a track record of acceptable internal and external validity supported by data provided by the database owners and peer-reviewed studies by external researchers. The General Practice Research Database and The Health Improvement Network are primary care UK databases that meet these criteria.[2–5]

COMPUTERISED SEARCH

The second step in the validation process is to establish a good operational definition of the outcome of interest by constructing specific diagnostic algorithms using a list of codes from the corresponding clinical dictionary. This initial computerised search might also include objective eligibility (exclusion) criteria. This list may be specific if we are dealing with some hard endpoints, such as cancer, but will have to be more sensitive (and will therefore have greater potential for false positives) when studying events such as peptic ulcer bleeding that may be coded with loose terms such as haematemesis or melaena.

MANUAL REVIEW OF COMPUTERISED PROFILES

The third step is a careful and, therefore, time-consuming manual review of computer profiles of the individuals identified by the algorithms in the previous step (computerised search) in order to assign a case status to each patient (probable, possible, or non-case). This first clinical review will permit the researcher to assess whether the validity of the initial computer search strategy is acceptable (confirmation rate close to 90%) or whether more information is required. If needed, the researcher will request from the database owner additional clinical information that the GPs include in a free-text section. This can include information derived from their narrative account of the episode, including summaries from referral letters, discharge letters, and diagnostic procedures.

The free-text section can sometimes also include a complete copy of these letters that have been scanned in. This will lead to a second clinical review of the whole list of individuals detected with the initial computer search, or a random sample if the authors simply wish to confirm a high correlation with the first clinical review (the one without free text). Needless to say, to perform this computer profile review successfully, well-trained experts in the disease field of interest, funding, and time are required: the process may take up to 1 year.

GP VALIDATION

The fourth step is obtaining confirmation of the case status from the GP. This is done by contacting collaborating GPs using specifically designed questionnaires. In addition to the questionnaire, a researcher can request — through the database owner —

anonymised copies of original medical records (consultant letters, hospital discharge, post-mortem reports). The final case validation with the GP is often done on a random sample of individuals when previous experience suggests that the positive predictive value will be high based on the case ascertainment obtained after the review of computerised profiles with or without free-text comments. Once again, time resources to complete this task are likely to extend from 4 months to 1 year depending on the number of questionnaires involved.

CASE VALIDATION EXAMPLES

In a recent study on acute urinary retention, we performed case validation without requesting a free-text section but with GP questionnaires.[6] We manually reviewed profiles of all patients with acute urinary retention (*n* = 4911) and classified them into probable or possible. After reviewing the returned questionnaires and all paper-based information related to the episode in a random sample, we confirmed a case of incident acute urinary retention in 92% of probable cases and 42% of possible cases. Given the degree of misclassification among possible cases, we used only probable acute urinary retention cases in our analyses. The time required to finalise this validation was close to 10 months.

Another example of the value of comprehensive validation of information recorded in CHCDs is a study on the risk of upper gastrointestinal complication among users of non-steroidal anti-inflammatory drugs.[7] The observed estimate of risk after performing complete review of patients with free-text comments and confirmation in a random sample with questionnaires sent to GPs was a relative risk of 3.7. The corresponding estimate using as cases all patients identified with the initial computer search was 2.6, a major underestimation of the effect.

Finally, a recent report evaluating different validation strategies showed the confirmation rate to increase from 76% (without free text) to 86% (incorporating free text) when classifying cases of hospitalised ischemic cerebrovascular events.[8] These reviews often result in changing the date of occurrence in a significant number of cases.[9]

Even though CHCDs are a revolutionary and rich information source for epidemiologic studies, as Khan *et al* suggested, the investigators conducting research using these databases need to consider carefully the limitations and strengths. Reducing the degree of outcome misclassification with CHCDs to an acceptable minimum is a methodological principle in epidemiologic research that has lost ground to other more fashionable developments trying to deal with confounding. Yet studies which control for exposure misclassification, selection bias, and confounding will still not provide valid findings if a case is not a case. Training to get the necessary expertise to use the information in these large databases is an important requisite for successfully performing studies with CHCDs. If researchers decide to skip this demanding validation process, then they should be ready to accept that journal reviewers and readers (when published) will have doubts about the validity and relevance of their findings. No-one has said that observational research using CHCDs is an easy endeavor. *Non sine pulvere palma.*

Luis Alberto García Rodríguez,
MD, MSc, Centro Español de Investigación Farmacoepidemiológica (CEIFE), Madrid, Spain.

Ana Ruigómez,
MD, PhD, Centro Español de Investigación Farmacoepidemiológica (CEIFE), Madrid, Spain.

Provenance
Commissioned; not peer reviewed.

Acknowledgements
We thank Drs Sonia Hernández Díaz and Miguel Gil who provided valuable comments to earlier versions of the manuscript, and Elisa Martín Merino and Lucía Cea Soriano for their valuable help in many validation exercises.

REFERENCES

1. Khan NF. Harrison SE, Rose PW. Validity of diagnostic coding within the General Practice Research Database: a systematic review. *Br J Gen Pract* 2010; **60(572):** 10.3399/bjgp10X483562.
2. García Rodríguez LA, Perez Gutthann S. Use of the UK General Practice Research Database for pharmacoepidemiology. *Br J Clin Pharmacol* 1998; **45(5):** 419–425.
3. Jick SS, Kaye JA, Vasilakis-Scaramozza C, *et al*. Validity of the General Practice Research Database. *Pharmacotherapy* 2003; **23(5):** 686–689.
4. Lewis JD, Schinnar R, Bilker WB, *et al*. Validation studies of the health improvement network (THIN) database for pharmacoepidemiology research. *Pharmacoepidemiol Drug Saf* 2007; **16(4):** 393–401.
5. Lo Re V 3rd, Haynes K, Forde KA, *et al*. Validity of The Health Improvement Network (THIN) for epidemiologic studies of hepatitis C virus infection. *Pharmacoepidemiol Drug Saf* 2009; **18(9):** 807–814.
6. Martín-Merino E, García Rodríguez LA, Massó-González EL, Roehrborn CG. Do oral antimuscarinic drugs carry an increased risk of acute urinary retention? *J Urol* 2009; **182(4):** 1442–1448.
7. García Rodríguez LA, Barreales Tolosa L. Risk of upper gastrointestinal complications among users of traditional NSAIDs and COXIBs in the general population. *Gastroenterology* 2007; **132(2):** 498–506.
8. Ruigómez A, Martín-Merino E, García Rodríguez LA. Validation of ischemic cerebrovascular diagnoses in the health improvement network (THIN). *Pharmacoepidemiol Drug Saf* 2010 Feb 3. [Epub ahead of print].
9. Margulis AV, García Rodríguez LA, Hernández-Díaz S. Positive predictive value of computerized medical records for uncomplicated and complicated upper gastrointestinal ulcer. *Pharmacoepidemiol Drug Saf* 2009; **18(10):** 900–909.

10.3399/bjgp10X483472

ADDRESS FOR CORRESPONDENCE

Luis Alberto García Rodríguez
Centro Español de Investigación Farmacoepidemiológica (CEIFE), Almirante 28 (2º), 28004 Madrid, Spain.
E-mail: lagarcia@ceife.es