Sarah Price

The Clinical Practice Research Datalink: How much symptom information is lost in the free text

Abstract

Databases of electronic medical records – such as the Clinical Practice Research Datalink (CPRD) – are increasingly used in epidemiological research. The outputs of such research may inform the evidence base underpinning national guidelines, whose recommendations influence large amounts of healthcare spending. Researchers using CPRD data have complete access to everything that GPs record using codes; however, any information they note in text boxes is not routinely available, primarily because its content may identify the patient. The CPRD will provide access to extracts of the text record, in line with a researcher's specified search terms, although this service is fairly costly owing to the extensive manual checks required to ensure complete anonymisation of the data. Therefore, the default is for researchers to restrict their analysis to just the coded data and to make the assumption that this is a representative sample of the complete medical record. Sarah has investigated whether omission of these 'hidden' text records introduces bias, and will report on her findings relating to the risk of cancer in symptomatic patients presenting to primary care.