Using routine health data for research: the devil is in the detail

Hannah Whittaker,[1] Jennifer K Quint [2]

Electronic healthcare records (EHRs) are increasingly being used for population-based studies globally. Despite their strengths, hidden pitfalls exist and researchers must take extra care to ensure high-quality data to minimise measurement error and biasses. This article discusses the recent work by Kerkhof *et al*, in relation to disease misdiagnosis and misclassification, the importance of linked data sources and the usability of test variables; all of which are extremely important issues that researchers must be aware of when using EHRs. The devil is in the detail.

EHR databases systematically and routinely collect and store healthcare data electronically and can include data on routine processes in primary and secondary care (disease codes, prescriptions, procedures and tests). The information collected ranges from medical insurance claims, to mortality data, to specific disease registries, with each database coding and storing information differently. The original purpose of EHRs was simply to store medical information digitally. But EHRs are increasingly being used for research and population-based studies globally, offering large sample sizes, a wide breadth of study variables and the inclusion of more generalisable populations.

However, nothing is ever perfect and routinely collected EHR data can have issues; the devil is in the detail. Unlike studies which include purposeful prospective data collection, the original purpose of data collection for EHRs is not research; a controversial argument in data science.[1] So, while EHRs allow researchers to study real-world populations seen in every day clinical practice, extra care must be taken to ensure the quality of the data is high in order to minimise measurement error and biasses.

In this issue of the journal, Kerkhof *et al* investigate whether acute exacerbations of chronic obstructive pulmonary disease (AECOPD) are associated with the rate of forced expiratory volume in 1 s (FEV_1) decline, depending on inhaled corticosteroid (ICS) use, in a UK COPD population. The authors additionally stratified blood eosinophil (EO) level to understand how EOs modify the relationship between AECOPD, ICS and the rate of FEV_1 decline. Patient's highest level of maintenance therapy (in order of long-acting β2-agonist (LABA), ICS, ICS/LABA, long-acting muscarinic antagonist (LAMA)/LABA, and LAMA/LABA/ICS) was determined and patients were grouped into ICS and non-ICS users.[2] AECOPDs were identified after the initiation of highest level of the maintenance therapy. Two large primary care EHRs were used: Clinical Practice Research Datalink (CPRD) and Optimum Patient Care Research Database. While the authors made efforts to comprehensively define the study population and exposure of interest, variables were sometimes lacking strength and definition due to disease misdiagnosis and misclassification, lack of linked data sources and the usability of test variables and their stability; a few pitfalls of using EHR data.

Disease misclassification in EHRs is extremely important to be aware of, especially when defining study populations. Using COPD and asthma as an example, a previous study found that 50% of COPD patients in CPRD had an asthma diagnosis ever recorded in their medical history; a large overestimation of the true prevalence of asthma in COPD.[3 4] Both COPD and asthma diagnoses have been validated separately in CPRD but more recently, ways in which concomitant diagnosis of COPD and asthma is identified in primary care has been studied.[5] Kerkhof *et al* excluded COPD patients with an asthma diagnosis on or after the first date of COPD diagnosis to exclude patients with current asthma or misdiagnosed COPD. A further sensitivity analysis excluded patients with an asthma diagnosis recorded at any point in their medical history in order to eliminate misclassification between COPD and asthma. While arguably a stringent approach, this may have excluded patients with (1) a history of asthma, including patients with childhood asthma, but who still have a valid COPD diagnosis, and (2) have an asthma diagnosis within 2 years of their COPD diagnosis thus excluding patients who likely had COPD, never asthma.

So, what about the use of linked data sources? AECOPD are a common study endpoint both in trials and EHR data. Identification of AECOPD have been validated in primary care and secondary care data.[6 7] By nature, secondary care events are more severe than AECOPDs treated in primary care. Therefore, if only primary care data are used, the frequency and severity of AECOPD events are underestimated.[7] While Kerkhof *et al* seemingly accurately identify AECOPD events in primary care, secondary care data was not used to define events. One could argue that hospitalisations should be recorded in primary care data; however, we know this is not always the case. Using only primary care data could have biassed results as not all AECOPD were detected. Additional use of secondary care data would have added value and could have provided further information on the associations seen.

Lastly, let's consider how laboratory values in EHR compare with those collected in a prospective setting. The authors identified single EO measurements used to stratify analyses by high and low EO levels. One of the issues here is that we don't always know why a blood test was done at a particular time. It may be that people who are sicker or having frequent healthcare more often are more likely to have a test done, leading to selection bias. In terms of identifying EOs, no validated primary care algorithms exist to date but previous studies have investigated the stability of EOs over time, which can be used to help provide definition.[8 9] These studies suggest that EOs measurements within 2 years are likely to remain similar and recommend a 2-year period for identification. To contextualise this, greater than 80% of COPD patients in UK EHR and US EHR cohorts had EOs <300 cells/μL in the first and second years of follow-up.[9] Kerkhof *et al* identified a single EO measurement over a 4-year period (2 years prior and 2 years after highest maintenance therapy initiation). It is highly likely that a single measurement taken over a 4-year period will not truly represent baseline EOs. Despite using this time window, the authors highlight that 82% of EO measurements were within 1 year of highest therapy initiation. As with other continuous variables, precision of recording of EOs is a major issue; one that agrees with the argument that data should only be used for the purpose it is collected.[1] For example, as Kerkhof *et*

[1]NHLI, Imperial College London, London, UK
[2]Respiratory Epidemiology, Occupational Medicine and Public Health, Imperial College London, London, UK

Correspondence to Dr Jennifer K Quint, Respiratory Epidemiology, Occupational Medicine and Public Health, Imperial College London, London SW7 2BU, UK; j.quint@imperial.ac.uk

al correctly highlighted, an EO recorded as 0.3×10^9/L could be an EO reading of anywhere between 250 and 349 cells/µL. Careful consideration of variable processing is important in ensuring high-quality data for research.

Despite the pitfalls, studies using EHRs are extremely important in adding to the literature commonly dominated by randomisedcontrolled trials (RCTs), so much so that the UK National Institute for Health and Care Excellence (NICE) and the US Food and Drug Administration (FDA) are trying to understand more and more on how to incorporate findings from EHR studies into guidelines. There is no doubt that RCTs are essential in medical research; however, given the specific populations of RCTs, real-world studies are needed to investigate research questions in more generalisable clinical settings. The more we use EHR in these types of studies, the more we can narrow down definitions and share validated definitions to strengthen the field. After all, the devil is in the detail.

Contributors HW and JKQ wrote and edited this work.

Funding The authors have not declared a specific grant for this research from any funding agency in the public, commercial or not-for-profit sectors.

Competing interests None declared.

Patient consent for publication Not required.

Provenance and peer review Commissioned; externally peer reviewed.

To cite Whittaker H, Quint JK. *Thorax* 2020;**75**:714–715.

Accepted 1 June 2020
Published Online First 24 June 2020

Linked

► http://dx.doi.org/10.1136/thoraxjnl-2019-214457

Thorax 2020;**75**:714–715.
doi:10.1136/thoraxjnl-2020-214821

ORCID iD
Jennifer K Quint http://orcid.org/0000-0003-0149-4869

REFERENCES

1 Peek N, Rodrigues PP. Three controversies in health data science. *Int J Data Sci Anal* 2018;6:261–9.
2 Kerkhof M, Voorham J, Dorinsky P, *et al*. Association between COPD exacerbations and lung function decline during maintenance therapy. *Thorax* 2020;75:744–53.
3 Quint JK, Müllerova H, DiSantostefano RL, *et al*. Validation of chronic obstructive pulmonary disease recording in the clinical practice research Datalink (CPRD-GOLD). *BMJ Open* 2014;4:e005540.
4 Nissen F, Morales DR, Mullerova H, *et al*. Validation of asthma recording in the clinical practice research Datalink (CPRD). *BMJ Open* 2017;7:e017474.
5 Nissen F, Morales DR, Mullerova H, *et al*. Concomitant diagnosis of asthma and COPD: a quantitative study in UK primary care. *Br J Gen Pract* 2018;68:e775–82.
6 Rothnie KJ, Müllerová H, Hurst JR, *et al*. Validation of the recording of acute exacerbations of COPD in UK primary care electronic healthcare records. *PLoS One* 2016;11:e0151357.
7 Rothnie KJ, Müllerová H, Thomas SL, *et al*. Recording of hospitalizations for acute exacerbations of COPD in UK electronic health care records. *Clin Epidemiol* 2016;8:771–82.
8 Landis SH, Suruki R, Hilton E, *et al*. Stability of blood eosinophil count in patients with COPD in the UK clinical practice research Datalink. *COPD* 2017;14:382–8.
9 Vogelmeier CF, Kostikas K, Fang J, *et al*. Evaluation of exacerbations and blood eosinophils in UK and US COPD populations. *Respir Res* 2019;20:178.