

# Using electronic health record data for clinical research: a quick guide

Sophie H Bots<sup>1</sup>, Rolf H H Groenwold<sup>2,3</sup> and Olaf M Dekkers<sup>2,4</sup>

<sup>1</sup>Division of Pharmacoepidemiology and Clinical Pharmacology, Utrecht Institute for Pharmaceutical Sciences, Utrecht, the Netherlands, <sup>2</sup>Department of Clinical Epidemiology, <sup>3</sup>Department of Biomedical Data Sciences, and <sup>4</sup>Department of Endocrinology, Leiden University Medical Center, Leiden, the Netherlands

Correspondence  
should be addressed  
to O M Dekkers  
**Email**  
o.m.dekkers@lumc.nl

## Abstract

Electronic health record (EHR) data not only offer many exciting research opportunities but also come with their own inherent limitations. Researchers may not always realise the challenges associated with the use of EHR data for research, or the fact that using large datasets of ‘real-world data’ does not necessarily provide valuable real-world evidence. This article discusses some of the main differences between EHR data and data collected primarily for research purposes, and the challenges encountered when using EHR data for research. It also offers suggestions on how to deal with these challenges based on worked-out examples. It therefore serves as a quick guide for researchers interested in either reading or performing EHR-based research.

*European Journal of  
Endocrinology*  
(2022) **186**, E1–E6

## Introduction

The digitalisation of the healthcare system has made a wealth of data collected during clinical care available for researchers. These electronic health record (EHR) data reflect the patient population seen in regular care and thus include patients who are often underrepresented in clinical trials, such as women, the elderly, and those with multiple co-morbidities (1, 2, 3). In addition, EHR data usually combine a relatively large sample size with a broad range of measurements and clinical outcome data. An example is a study on long-term clinical outcomes in Cushing’s disease (4), where the outcomes of interest (mortality, infections, thrombosis) were routinely captured in registries with nationwide coverage. This is much harder to accomplish when collecting data in a ‘traditional’ way, for example, as part of a research cohort, due to the limited availability of time and other resources.

Large samples that are representative of daily clinical practice allow us – in theory – to answer clinically relevant questions even for rare diseases or heterogeneous patient populations. However, findings from EHR studies are not automatically more likely to be true just because they are

based on large study populations and data from routine daily care. In fact, large sample sizes generally do not affect underlying biases. They do however increase the precision of statistical tests, meaning that biased results are more likely to become statistically significant (5). In addition, EHR data are collected for various reasons, including clinical practice support, reimbursement purposes, and comparisons between healthcare providers. As a consequence, EHR data are inherently different from data collected specifically for research purposes (6, 7, 8). Ignoring these inherent differences could negatively impact both the feasibility and the validity of any research project using EHR data. The considerations below serve as a quick guide for designing or reading a study based on EHR data.

## Fitness for purpose

For a dataset to be fit for purpose for research, should contain all the information needed to answer the

research question (6). While this may seem obvious, it is fundamental to check whether the variables of interest are properly measured within the EHR. This concerns both whether these variables are (potentially) available (scope) (6, 8, 9, 10), and whether they are measured well enough to be useful (quality) (6, 8, 9).

As an example, take the research question of whether acromegaly increases the risk of heart failure (HF) and consider if EHR data would be fit for the purpose to answer this question. First off, HF is a tricky case. Multiple diagnostic approaches for HF have been developed and used over time or even concurrently, and some of these approaches may require specific echocardiographic or laboratory measurements that are not always part of daily care. The absence of such measurements will limit the scope of the EHR dataset, as this means that recorded HF diagnoses may be substandard. In addition, some of these parameters may be difficult to measure correctly or have large inter-observer variability. This may negatively affect the quality of the EHR dataset as a diagnostic HF code may have different meanings depending on who entered it. If both the scope and quality of the HF diagnosis are affected to such an extent that the outcome of HF cannot be reliably ascertained, the EHR dataset could be considered not fit for purpose. Mind that this problem is not solved by studying hospital admission for HF instead of HF *per se*, as HF admissions may also be prone to misclassification, or different classifications may have been used over time. While HF misclassification is a limitation for the acromegaly research question, it may be less problematic for other questions. For instance, time trends in HF diagnosis can be studied as long as the misclassification is the same across the studied time period. Although in that case the incidence rates may be incorrect, the observed fluctuation in HF diagnoses over time may reflect the true variation in HF diagnoses provided no drastic changes in surveillance or diagnostic criteria have occurred during the time period of interest. This illustrates how the same EHR data can be fit for purpose for some questions, but not others.

Another common situation is that information is potentially available but is either hidden in free text fields or only obtainable via linkage to external registries. HF diagnosis could be recorded in a free text field such as the general practitioner summary letter and this could be used to identify HF patients. This approach requires text retrieval methods to turn free text into an analysable format and needs a quality check of both the text fields and the chosen retrieval method. The retrieval method must balance identifying true HF diagnoses against obtaining incorrect (i.e. false positive) diagnoses where, for

example, the diagnosis was only considered but ultimately not confirmed.

When the required information is not recorded in a certain dataset or registry, linkage with other registries could be a solution. An example would be linking a hospital discharge database with a pharmacy registry to identify patient subgroups based on medication use, such as differentiating between type 1 and type 2 diabetes. However, patients may not be represented in both registries or the available unique identifiers may not be sufficient for a perfect match, for example, if only age and postal code are available. This could impact the quality of the linked dataset. It must also be noted that not all EHR datasets may be eligible for linkage, for example, if this threatens to de-anonymise the dataset or if probabilistic linkage is required and the percentage of mismatches is expected to be large.

## Primary sources of bias in EHR data

If at first glance the scope and quality of an EHR dataset appear fit for purpose, it is important to consider other types of bias that may be introduced by specific aspects of EHR data. Although these biases may also affect more traditional research, they are omnipresent in research using EHR data. Common sources of bias in EHR data are discussed below.

### Selection bias

Selection bias is a form of bias that is generally introduced by a selective process of inclusion of subjects into a study or into the data analysis of a study (11). Continuing our HF example, one of the diagnostic rules may require an NT-proBNP biomarker measurement to differentiate HF from other conditions with similar symptoms such as chronic obstructive pulmonary disease. Guidelines recommend measuring it in patients suspected of HF (12), which already implies that NT-proBNP is not routinely measured in all patients. If the population of interest is all HF patients, selecting only those with an NT-proBNP measurement introduces selection bias because patients with symptoms suggestive of HF are more likely to be included than those with less typical symptoms.

It is also important to consider that patients generally only enter an EHR database after contact with the medical health care system and that this can occur anywhere in the patient's clinical trajectory. As a result, some individuals may never enter an EHR database, and similarly, patients

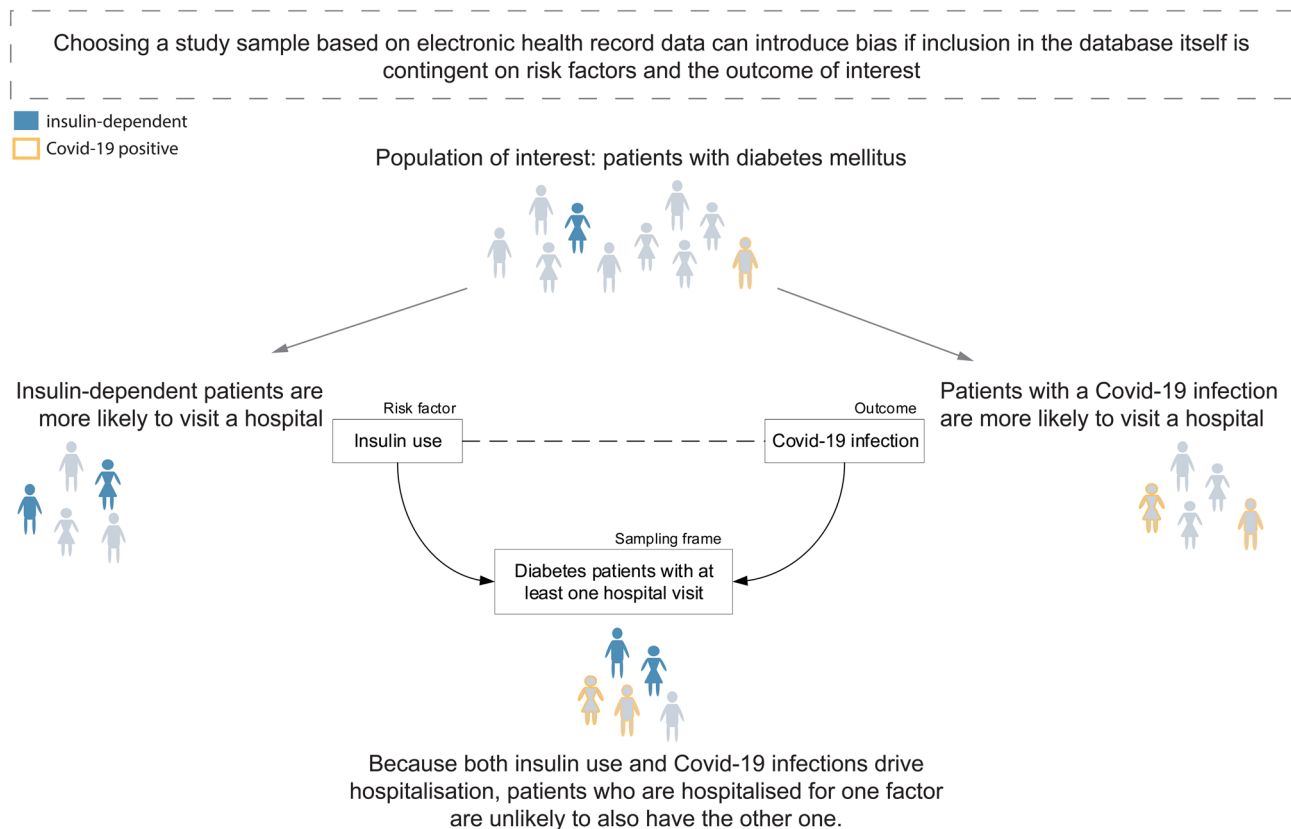
with mild cases of a disease may be absent from a hospital EHR database. This can induce a specific type of selection bias (known as collider stratification bias (13)) when both risk factors and the outcome of interest affect the likelihood to be included in an EHR database.

For example, researchers may want to know whether insulin use in type 2 diabetes patients increases the risk of Covid-19. To answer this question, they use hospital-based EHR data. In this situation, the likelihood that information about a patient with diabetes is available in hospital-based EHR data is influenced by both diabetes-related factors (i.e., insulin use) and the severity of Covid-19. Diabetes patients who require insulin may be more likely to be hospitalised than those who do not (as insulin use is a marker of disease stage in diabetes type 2), and diabetes patients with a Covid-19 infection are also more likely to be hospitalised than those without Covid-19 (having two diseases increases the risk for hospitalisation compared to having one disease only). In this group of hospitalised diabetes patients, those who do not use insulin are more likely to have Covid-19 than those who do use insulin. Likewise, those who do not have Covid-19 are more likely

to use insulin than those who do have Covid-19 because in both instances there must have been a reason why those patients were hospitalised. Thus, even if there is no true relation between insulin use and Covid-19, a spurious (inverse) association is observed in the dataset due to the selection process. If such bias is not recognised, researchers may falsely conclude that insulin use prevents against Covid-19. This is illustrated in Fig. 1 (for more details see (13)).

## Confounding

Another source of bias in studies using EHR data is the inherent incomparability between those who are treated (or exposed) and those who are not, which is generally referred to as confounding. Consider a study of the effect of growth hormone lowering medication that compares treated and untreated acromegaly patients with a mild recurring disease. Incomparability in measured patient characteristics can be controlled for in the analyses, but this is impossible for any reasons to start, stop, or withhold treatment that were not measured. The potential



**Figure 1**

An example of collider stratification bias.

for such unmeasured confounding is arguably larger in EHR data because the quality is often lower than that of more traditional research data. Note that the actual magnitude of unmeasured confounding likely depends on the comparison that is being made (14). In addition, confounding only plays a role in studies that aim to investigate causal effects. Confounding is no issue in types of research that do not make causal claims, for example diagnostic test accuracy research.

### Information bias, missing data and misclassification

Information bias may occur when the information used in a study does not (perfectly) represent the phenomenon that is studied, for example because exposure or outcome variables are measured with an error or when categorical data are misclassified. This has been discussed in detail previously (15, 16), but there are a few considerations particularly relevant for EHR data.

First, there is a type of missingness that is inherent to EHR data called informative missingness, where the absence (or presence) of information holds information in itself (17). Informative missingness occurs for instance when a healthcare professional does not record certain information on purpose because these are not relevant for that patient. For example, growth hormone levels are usually not measured in patients with type 2 diabetes mellitus unless the treating physician expects growth hormone to play a pivotal role in the development of diabetes in that patient. Thus, the absence of such a measurement indicates that the growth hormone level was likely normal. Yet if a growth hormone measurement was performed in a patient with diabetes, there must have been an underlying reason, and thus the fact that there is a measurement also holds clinical information regarding the patient's health. In the case of diagnoses or prescriptions of medication, the absence of a recording can reasonably be assumed to imply that the diagnosis or prescription was absent. In such cases, a missing data entry can be filled in ('imputed') by a value representing absence (usually the value zero). However, for continuous variables like growth hormone levels, this approach is not valid and is more sophisticated methods to impute missing data such as multiple imputations should be considered. However, even these methods may not be a solution in the case of informative missingness (18). Importantly, methods to mitigate missing data depend on the aims of the study (19) and there is no universally perfect strategy to deal with

missing data in research using EHR data. Researchers thus need to make their assumptions about missingness patterns in their data explicit, apply the missing data strategy they deem most appropriate and provide a rationale for the statistical choices made.

Second, data entry errors are common in EHR data because there is no routinely implemented quality check and many different healthcare professionals enter data. In addition, data collection and data entry are not standardised, leaving professionals with some flexibility to do what they think is best within what is recommended by guidelines. For example, some may choose to always measure blood pressure twice regardless of the value, while others may only repeat the measurement when the first value is above the normal upper limit. More extremely, certain patient groups may be monitored more closely than others and thus have a higher chance of having conditions they suffer from being diagnosed, a type of information bias known as detection (or surveillance or ascertainment) bias. Going back to the NT-proBNP example, using a definition of HF that requires an NT-proBNP measurement can introduce bias if certain patient groups have a higher or lower likelihood of undergoing this measurement than others. For example, researchers may want to compare the risk of HF (defined based on NT-proBNP) in acromegaly patients against the risk in the general population. Suppose, patients with acromegaly have their NT-proBNP measured as part of their standard workup, whereas patients in the general population only undergo this measurement in case of a clinical suspicion of HF. In this situation, milder cases of HF are more likely to be picked up in acromegaly patients than in the general population, which biases the comparison between the two groups.

This flexibility may also lead to measurement errors and misclassification (and thus bias), and also complicates generalisation between (EHR) datasets because of variation in data collection methods (20). Moreover, many routine codes used in clinical care have an inherent degree of misclassification, of which the ICD-10 codes are an example (21). It is thus recommended to familiarise oneself with the coding system being used and provide the rationale for choosing certain diagnostic codes to define study outcomes or exposures, also because coding systems can vary across EHR datasets. Ideally, data quality is checked before analysis to filter out data entry errors or identify other quality issues. This could be done by detailed quality assessment of subsets of the data, a so-called subset validation study.

## Conclusion

EHR data offer a wealth of opportunities but also come with their own set of inherent limitations. This quick guide is not an exhaustive list of all opportunities and limitations and for example does not mention time investment needed to operationalise raw EHR data for research, but it has discussed the main methodological and statistical points researchers should consider when performing or reading EHR-based research. Most importantly, the reasons behind data collection differ between traditional research data, generally collected to specifically answer certain research questions, and EHR data, collected driven by medical need. This can affect both how fit for purpose an EHR dataset is and which sources of bias could be present, depending on the research question of interest. It is therefore important to realise that using 'real-world evidence' from EHR datasets does not automatically provide valuable real-world evidence. Indeed, due to the challenges associated with using EHR data, the validity of methodological and statistical choices in real-world studies is never guaranteed and should be substantiated by the researchers who use and are familiar with the intricacies of EHR data. A pre-specified analysis protocol may help in making potential pitfalls explicit and should be considered for any EHR-based study.

### Declaration of interest

The authors declare that there is no conflict of interest that could be perceived as prejudicing the impartiality of this editorial. O M Dekkers is a Deputy Editor for *European Journal of Endocrinology*. O M Dekkers was not involved in the peer review or editorial process for this paper, on which he is listed as an author.

### Funding

This work did not receive any specific grant from any funding agency in the public, commercial or not-for-profit sector.

## References

- Pilote L & Raparelli V. Participation of women in clinical trials: not yet time to rest on our laurels. *Journal of the American College of Cardiology* 2018 **71** 1970–1972. (<https://doi.org/10.1016/j.jacc.2018.02.069>)
- Sardar MR, Badri M, Prince CT, Seltzer J & Kowey PR. Underrepresentation of women, elderly patients, and racial minorities in the randomized trials used for cardiovascular guidelines. *JAMA Internal Medicine* 2014 **174** 1868–1870. (<https://doi.org/10.1001/jamainternmed.2014.4758>)
- Van Spall HG, Toren A, Kiss A & Fowler RA. Eligibility criteria of randomized controlled trials published in high-impact general medical journals: a systematic sampling review. *JAMA* 2007 **297** 1233–1240. (<https://doi.org/10.1001/jama.297.11.1233>)
- Dekkers OM, Horváth-Puhó E, Jørgensen JO, Cannegieter SC, Ehrenstein V, Vandenbroucke JP, Pereira AM & Sørensen HT. Multisystem morbidity and mortality in Cushing's syndrome: a cohort study. *Journal of Clinical Endocrinology and Metabolism* 2013 **98** 2277–2284. (<https://doi.org/10.1210/jc.2012-3582>)
- Meng X-L. Statistical paradises and paradoxes in big data (I): law of large populations, big data paradox, and the 2016 US Presidential Election. *Annals of Applied Statistics* 2018 **12** 685–726. (<https://doi.org/10.1214/18-AOAS1161SF>)
- Verheij RA, Curcin V, Delaney BC & McGilchrist MM. Possible sources of bias in primary care electronic health record data use and reuse. *Journal of Medical Internet Research* 2018 **20** e185. (<https://doi.org/10.2196/jmir.9134>)
- Weiskopf NG & Weng C. Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. *Journal of the American Medical Informatics Association* 2013 **20** 144–151. (<https://doi.org/10.1136/amiajnl-2011-000681>)
- Hersh WR, Weiner MG, Embi PJ, Logan JR, Payne PR, Bernstam EV, Lehmann HP, Hripcsak G, Hartzog TH, Cimino JJ *et al.* Caveats for the use of operational electronic health record data in comparative effectiveness research. *Medical Care* 2013 **51** (Supplement 3) S30–S37. (<https://doi.org/10.1097/MLR.0b013e31829b1dbd>)
- Overhage JM & Overhage LM. Sensible use of observational clinical data. *Statistical Methods in Medical Research* 2013 **22** 7–13. (<https://doi.org/10.1177/0962280211403598>)
- Weiner MG & Embi PJ. Toward reuse of clinical data for research and quality improvement: the end of the beginning? *Annals of Internal Medicine* 2009 **151** 359–360. (<https://doi.org/10.7326/0003-4819-151-5-200909010-00141>)
- Hernán MA, Hernández-Díaz S & Robins JM. A structural approach to selection bias. *Epidemiology* 2004 **15** 615–625. (<https://doi.org/10.1097/01.ede.0000135174.63482.43>)
- Ponikowski P, Voors AA, Anker SD, Bueno H, Cleland JGF, Coats AJS, Falk V, González-Juanatey JR, Harjola VP, Jankowska EA *et al.* 2016 ESC guidelines for the diagnosis and treatment of acute and chronic heart failure. *Revista Espanola de Cardiologia* 2016 **69** 1167. (<https://doi.org/10.1016/j.rec.2016.11.005>)
- Griffith GJ, Morris TT, Tudball MJ, Herbert A, Mancano G, Pike L, Sharp GC, Sterne J, Palmer TM, Smith GD *et al.* Collider bias undermines our understanding of COVID-19 disease risk and severity. *Nature Communications* 2020 **11** 5749. (<https://doi.org/10.1038/s41467-020-19478-2>)
- Houwert RM, Beks RB, Dijkgraaf MGW, Roes KCB, Öner FC, Hietbrink F, Leenen LPH & Groenwold RHH. Study methodology in trauma care: towards question-based study designs. *European Journal of Trauma and Emergency Surgery* 2021 **47** 479–484. (<https://doi.org/10.1007/s00068-019-01248-5>)
- Groenwold RHH & Dekkers OM. Missing data: the impact of what is not there. *European Journal of Endocrinology* 2020 **183** E7–E9. (<https://doi.org/10.1530/EJE-20-0732>)
- Groenwold RHH & Dekkers OM. Measurement error in clinical research, yes it matters. *European Journal of Endocrinology* 2020 **183** E3–E5. (<https://doi.org/10.1530/EJE-20-0550>)
- Haneuse S, Arterburn D & Daniels MJ. Assessing missing data assumptions in EHR-based studies: a complex and underappreciated task. *JAMA Network Open* 2021 **4** e210184. (<https://doi.org/10.1001/jamanetworkopen.2021.0184>)
- Groenwold RHH. Informative missingness in electronic health record systems: the curse of knowing. *Diagnostic and Prognostic Research* 2020 **4** 8. (<https://doi.org/10.1186/s41512-020-00077-0>)
- Sperrin M, Martin GP, Sisk R & Peek N. Missing data should be handled differently for prediction than for description or causal explanation. *Journal of Clinical Epidemiology* 2020 **125** 183–187. (<https://doi.org/10.1016/j.jclinepi.2020.03.028>)



20 Luijken K, Wynants L, van Smeden M, Van Calster B, Steyerberg EW, Groenwold RHH & Collaborators. Changing predictor measurement procedures affected the performance of prediction models in clinical examples. *Journal of Clinical Epidemiology* 2020 **119** 7–18. (<https://doi.org/10.1016/j.jclinepi.2019.11.001>)

21 Harriss LR, Ajani AE, Hunt D, Shaw J, Chambers B, Dewey H, Frayne J, Beauchamp A, Duvé K, Giles GG *et al.* Accuracy of national mortality codes in identifying adjudicated cardiovascular deaths. *Australian and New Zealand Journal of Public Health* 2011 **35** 466–476. (<https://doi.org/10.1111/j.1753-6405.2011.00739.x>)

---

Received 25 October 2021

Revised version received 24 January 2022

Accepted 14 February 2022