OPEN ACCESS

From real-world electronic health record data to real-world results using artificial intelligence

Rachel Knevel [1,2] Katherine P Liao[3,4]

Handling editor Josef S Smolen

[1]Department of Rheumatology, Leiden University Medical Center, Leiden, The Netherlands
[2]Newcastle University School of Clinical Medical Sciences, Newcastle upon Tyne, UK
[3]Division of Rheumatology, Immunology, and Allergy, Brigham and Women's Hospital, Boston, Massachusetts, USA
[4]Harvard Medical School Center for Biomedical Informatics, Boston, Massachusetts, USA

Correspondence to
Dr Rachel Knevel, Department of Rheumatology, Leiden University Medical Center, Leiden 2333 ZA, The Netherlands; r.knevel@lumc.nl

Received 15 July 2022
Accepted 10 September 2022
Published Online First 23 September 2022

To cite: Knevel R, Liao KP. *Ann Rheum Dis* 2023;**82**:306–311.

ABSTRACT

With the worldwide digitalisation of medical records, electronic health records (EHRs) have become an increasingly important source of real-world data (RWD). RWD can complement traditional study designs because it captures almost the complete variety of patients, leading to more generalisable results. For rheumatology, these data are particularly interesting as our diseases are uncommon and often take years to develop. In this review, we discuss the following concepts related to the use of EHR for research and considerations for translation into clinical care: EHR data contain a broad collection of healthcare data covering the multitude of real-life patients and the healthcare processes related to their care. Machine learning (ML) is a powerful method that allows us to leverage a large amount of heterogeneous clinical data for clinical algorithms, but requires extensive training, testing, and validation. Patterns discovered in EHR data using ML are applicable to real life settings, however, are also prone to capturing the local EHR structure and limiting generalisability outside the EHR(s) from which they were developed. Population studies on EHR necessitates knowledge on the factors influencing the data available in the EHR to circumvent biases, for example, access to medical care, insurance status. In summary, EHR data represent a rapidly growing and key resource for real-world studies. However, transforming RWD EHR data for research and for real-world evidence using ML requires knowledge of the EHR system and their differences from existing observational data to ensure that studies incorporate rigorous methods that acknowledge or address factors such as access to care, noise in the data, missingness and indication bias.

BACKGROUND

Real-world data (RWD) is defined as 'data relating to patient health status and/or the delivery of healthcare routinely collected from a variety of sources'.[1] While there are several types of RWD, such as claims data and patient registries, the use of electronic health record (EHR) data for clinical studies is perhaps the fastest growing segment. This growth can be attributed to several factors, including the increasing adoption of EHRs[2] and digital technologies that register healthcare processes stored in EHRs. In large EHRs, millions of data points are available in millions of patients, reflecting myriad patient paths through the medical system. However, extracting generalisable knowledge from RWD is challenging due to issues that arise from any dataset not designed for research such as confounding, missingness and heterogeneity in how the data are documented, for example, clinical notes. Fortunately, growing in parallel to the increased ability to measure and capture health related data, were advances in computing to store and process, and methods to analyse these data, notably artificial intelligence (AI). Thus, the combination of rich clinical data available in EHRs, paired with the ability to analyse these data with AI have expanded the opportunities to better understand the diseases and people whom we treat.

Rheumatology research particularly benefits from studies using EHR data. Rheumatic conditions are generally uncommon. To enrol sufficient numbers of patients for population-based studies requires years to decades. The majority of rheumatic diseases are also chronic, and benefit from datasets where patients are followed longitudinally. EHRs with their existing large populations enable the potential to study the majority of subjects with a rheumatic condition followed in the healthcare system without requiring in-person recruitment. In addition, the patients' digital health records capture multiple health domains, for example, clinical notes, vital sign data, laboratory measurement, drug prescriptions, over time providing the opportunity to examine and generate new insights into disease progression, risk factors and management

AI and particularly machine learning (ML) methods, a subset of AI, have been particularly useful in their ability to handle the volume and heterogeneity of RWD. As RWD and AI become increasingly incorporated into studies and clinical care, knowledge of the strengths and limitations will become increasingly important for all medical specialists. This review will focus on the opportunities and challenges of using RWD focused mainly on EHR data, to advance clinical research in rheumatology, and where we may translate the methods and findings into clinical practice.

RWD-EHR expands the clinical data available to address clinical research questions

There are two broad types of clinical data for research: observational data, which includes prospective cohort studies and RWD/EHR, and clinical trials (table 1). In the hierarchy of clinical evidence, randomised controlled trials (RCTs) sit at the top largely because they are less prone to bias compared with other available datasets. However, the RCT study design restricts the types of questions one can answer. Clinical trials are designed to test the effect of a particular intervention, for example, drug, surgery, on an outcome, for example, mortality, myocardial infarction. Clinical trials have strict inclusion criteria excluding patients with comorbidities and particular age groups. Homogenising the patient population facilitates clear comparison of the effect of the intervention

Table 1 Types of clinical data available for research studies

Characteristics	Observational data		
	RWD-EHR	Prospective longitudinal cohort study or registry	Clinical trial
Definition	Data from EHR relating to patient health status and/or the delivery of healthcare routinely collected from a variety of sources	Non-interventional clinical study, prospectively collecting data on a group of patients with a particular disease or symptom	Patients assigned to one or more interventions to evaluate its impact on healthcare outcomes, for example, randomised controlled trial
Patient population	Broad, encompassing medical system or population area	Restricted by study participation	Restricted eligibility criteria, often excluding elderly and people with comorbidities
Data types	High dimensional	High no, limited by research design and variables for collection decided a priori	Variables/outcomes for collection decided a priori
	Data collected as part of patient care from both patients and physicians	Structured data collection and questionnaires	
Data presence	Sparce, noisy	Structured, same data collected on all participants	Highly structured and often with detailed clinical data
	Missingness not at random	Fairly complete	Low missingness
Scale	Large, thousands to millions	Modest, hundreds to thousands	Small, tens to thousands
Generalisability	Strong local structure can restrict generalisability Incorporating real-life noise into the analyses improves applicability to real life settings	Easily replicable in similar designed cohorts Generalisability restricted by patient selection and data not always directly implementable to real life settings.	The more restrained the patient selection the less generalisable

EHR, electronic health record; RWD, real-world data.

but reduce generalisability of the findings to the true patient population. One example is with the paucity of women in preclinical studies of cardiovascular drugs. Studies mainly included men due to concern that the hormonal changes in women could influence the effectiveness of the drug. However, since women were excluded or less preferentially recruited, results from these studies lack generalisability to the 40%–60% of the true patient population.[3] Moreover, RCTs are often powered to answer one question on the main treatment effect, and are underpowered to determine if subgroups of patients may benefit from one treatment versus another. Importantly, the RCT study design is suboptimal to study other important aspects of diseases, including disease development and pathogenesis. For studies related to patient subgroups or disease development, larger cohorts are needed, where variation in the patient population is a strength, rather than a weakness.

Observational data include the majority of clinical data for research and include longitudinal prospective cohort studies, registries and RWD/EHR. Longitudinal prospective cohort studies were designed to study risk factors for and development of diseases. A well-known example is the Framingham Heart Study. Their data provided the basis for many of the cardiovascular risk estimators used in clinical care today.[4] Observational cohort studies are designed to follow patients with particular diseases, symptoms and/or exposures to observe how they evolve over time.[5] Observational cohorts take many years before all relevant information is collected, making it a fairly time and resource intensive process. To measure the disease progression or incidence of events, most cohorts have fixed visits, and a fixed set of clinical factors or outcomes for which patients are assessed, providing structure to the data. The cohorts have wider inclusion criteria and patients are generally more willing to participate as there is no trial intervention. While fixed visits with near complete data capture is an advantage, one pitfall of fixed visits is that they fail to capture the disease events in between the visits and retrospective questionnaires suffer from recall bias.[6] Finally, the type of the measurements taken, both in clinical trials and observational cohorts, are driven by researchers' hypothesis and decided on a priori, whereby not considered important initially can be missed.

RWD offers alternatives for the above-mentioned shortcomings in traditional study designs: it is generally more inclusive than observational cohort studies and RCTs, extensive, available and big. For these reasons, many studies, including RCTs, now leverage RWD to extend their data collection.[7] In this review, we focus on the use of a major type of RWD, EHR data.

Opportunities for RWD-EHR to catalyze science and healthcare within rheumatology

EHRs contain data as part of routine care, including unscheduled visits during a flare or hospitalizations, and can fill in data gaps not available from RCT and observational cohort studies. A key question in rheumatic conditions is evolution of the clinical history before and after onset of the condition.[8] A challenge for prospective patient collections is to capture patients at the right moment, particularly early in the disease. The low prevalence of autoimmune conditions and uncertainty about the initial symptoms is a barrier for creating cohorts that capture the true beginning of the loss of self-tolerance. RWD-EHR allows us to look back at previously collected data. RWD-EHR have led to findings such as the association between EBV exposure and multiple sclerosis development in the US military data, autoantibodies preceding SLE, as well as lifestyle and SLE development.[8–11] RWD-EHR can also capture data surrounding the time disease development compared with studies with fixed visits of trials and cohorts. In addition, the number of dimensions or types of data measured in the real world tends to be higher than RCT or observational cohort studies; EHR data contains all data collected as part of clinical care on all patients who visited a clinic or healthcare system. Thus, RWD-EHR generally contains a broader range of demographics, for example, age, sex or socioeconomic status compared with existing clinical datasets. Routine clinical care recorded in most EHR included detailed diagnoses codes, symptom description, disease development, treatment and comorbidities. This creates a dataset where associations between diseases and comorbidities can be identified which might have not been captured in the predesigned data collections. For instance, studying the association between checkpoint inhibitors and the diverse manifestations for immune related adverse events would have been difficult to design a priori.[12] Particularly for complex autoimmune diseases, where both the risk factors and the disease classifications are uncertain, the high dimensional EHR data allows for wide data exploration to detect unknown patterns.

RWD/EHR is complementary to traditional clinical datasets, as it provides information that is difficult to obtain otherwise. Inevitably, RWD has its own shortcomings: the data collections are less well structured, sparse and noisy and the missingness

is not at random, but informed by clinical decision-making. Handling EHR requires special attention to data selection and data analytics. With standing biobanks, the limitation is no longer recruiting and collecting samples for typing. The main limitation is now accurate phenotyping and subsequently to extract reliable novel knowledge. We will address these challenges and solutions in the upcoming paragraphs.

Transforming RWD-EHR data to research ready data, starting with phenotypes

Phenotypes are the foundation for clinical research. A major contribution of RWD-EHR data to rheumatic disease research is the ability to efficiently create large cohorts of uncommon conditions for studies. There are two main types of EHR data—structured, for example, diagnosis codes, electronic prescriptions and unstructured data, for example, narrative text notes, imaging data. Classifying rheumatic and autoimmune diseases can be challenging as the accuracy of diagnosis billing codes alone can be low, for example, RA with positive predictive value (PPV) ~20%.[13 14] In addition, for some, specific diagnosis codes did not exist, that is, acute CPP disease or pseudogout.[15] Since the majority of rheumatic conditions rely on clinical diagnoses, many of the key features important for diagnosis are often buried in the unstructured text notes, for example, synovitis, radiographic evidence of sacroiliitis. To mine the large and diverse data from EHR, AI has offered valuable solutions. ML, a subfield of AI, are computer systems that are able to independently learn and, ideally, generalise observed patterns from data. They are widely used for prediction and classification models. Since they can be developed using a high number of variables in large populations, they are very suited for building models that can be applied to the EHR to classify patients for inclusion into an EHR-based cohort. The same principles used to develop phenotype algorithms for research will also be used when developing algorithms for clinical care. Thus, we believe it is important for all healthcare providers to become familiar with the framework for how these algorithms are developed. Below, we review some of the key steps for consideration when building and evaluating a model for clinical phenotyping.

Model building for phenotyping

Perhaps the most important application of ML using EHR data is phenotyping: classifying patients with a disease and characterising patients.[16 17] Where clinical trials and prospective cohorts screen patients before inclusion, in RWD patients are selected retrospectively using the available data. The magnitude of EHR data makes chart review to classify all patients with a particular phenotype almost infeasible. Studies have found that relying on diagnostic or financial codes solely to create roust cohorts, is often not sufficiently precise[17 18] and classification models and ML techniques using a broader set of data from have served to fill that gap.[7]

Set gold standard

When setting the gold standard, the investigator is defining the phenotype that the algorithm will define, for example, 200 patients identified with psoriatic arthritis identified via chart review. However, in rheumatology the gold standard may not be as straightforward as defining for example diabetes or coronary arterie disease (CAD). First, there is still much discussion about what is true RA or SLE, with SLE-like and pre-RA disease types and several updates of the disease classification criteria. Second, since our diagnoses are based on the pattern recognition of multiple symptoms or abnormalities, a consensus defined set of diagnostic features may not be available. Clinicians are selective in what they record in the notes and thus checking for classification criteria, mainly designed for research studies, can result in an under-sampling of cases. Incorporating the final diagnosis of a rheumatologist as written might be more accurate as this captures the summary of the complete clinical reasoning and also factors that the rheumatologist did not record. Depending on the research question, the wider spectrum of phenotypes captured by the rheumatologist's diagnosis can be a particular reason to use RWD, instead of using the more narrow defined inclusion criteria of clinical studies.[19]

Feature selection

To build a phenotyping algorithm model, one can select variables or features based on clinical knowledge and hypotheses, or using a hypothesis-free approach using all available data. ML can learn patterns from a set of high dimensional training examples. It allows for a fast data processing of EHR combining both codified structured data, for example, lab results or treatment prescriptions in a fixed format, and unstructured data, free written text in clinical notes. To use the latter, natural language processing (NLP) can identify and synthesise structure in the (digital) clinical notes (for hand written one would first need to transform them to digital notes before applying NLP).[20] In rheumatology, NLP expands the previously difficult to access features for integration in the analysis, for example, bone erosions, seropositivity status, or the concept of a flare. The resulting features can be considered in the ML model. Most ML algorithms will provide probabilities of having a disease to each patient. When well calibrated, different thresholds can be used to create a more precise or more sensitive patient selection.

Phenotyping across EHR systems

Several phenotyping pipelines available online, some of which created in large consortia such as eMERGE and i2b2[13 20–23] have built highly accurate algorithms for phenotype selection, which are implemented in multiple centres. However, even these 'universal' algorithms require validation in each centre. When healthcare systems, EHR software and languages differ, such as in Europe, aiming for an universal algorithm is extremely challenging. For this, solutions are available to enable centres to build algorithms on their own data following an NLP ML pipeline.[20 21]

Supervised versus unsupervised learning for phenotyping

Most models for clinical studies rely on supervised learning. In supervised learning, the model is developed using gold standards defined by a clinical expert, for example, chart review containing 200 patients with and without psoriatic arthritis (PsA), with the goal to identify the pattern that exists between patients with PsA vs those without. Unsupervised ML models can also be used for this purpose when there is a need to phenotype multiple conditions.[24–26] These models are generally not as accurate as supervised models, but enable high-throughput phenotyping over a handful to thousands of phenotypes with improved accuracy over diagnoses codes alone. Moreover, unsupervised models have been used to build clinical models to predict disease courses, optimise diagnostics and target treatment.[27] Unsupervised pattern recognition analyses identify subgroups of patient-patient similarity in a high dimensional or graph-based space. In rheumatology, they are most commonly employed for biological studies for instance to differentiate cell types in high-dimensional typing of blood and synovial biopsies, and are increasingly applied to clinical data from observational studies and post-hoc analyses of

clinical trials.[28–30] The identification of homogeneous disease subsets and trajectories within these large datasets can support research to disease aetiology and optimise treatment, particularly in the setting of complex heterogeneous diseases. Whether a model is trained in a supervised or unsupervised manner, accurate and generalisable results are important. For this, there are analytical steps important in ML.

Measuring performance of supervised models

Since ML aims to classify and predict, the key performance features are AUC-ROC (tradeoff between sensitivity and specificity), area under the precision recall curve (AUC-PRC) (trade-off between sensitivity and PPV) and F1-score (harmonic mean of sensitivity and PPV),[31–33] in addition to assessing the calibration (whether the magnitude of the probabilities (low, intermediate or high) are consistently accurate. When a probability threshold is set, the accuracy of predictions can be expressed by sensitivity, specificity, PPV and negative predictive value. Finally, the impact of a model's measurements can be calculated with net benefit and numbers needed to threat.

Developing balanced and reliable algorithms

To prevent overfitting a commonly applied technique for model optimisation and validation, is to divide the original data into a training-out and a hold-out test set, ideally in an iterative way such as in *k*-fold cross validation or leave-one-out cross-validation. The performance of the final model is then summarised by taking the average performance across all iterations for a robust assessment. Once the model is set, the final test round should ideally be done in data that was not used in any of the previous stages. The model's performance in the final round is considered the true performance, that is, internal validation. When assessing the validity and usefulness of an algorithm, it is imperative to check the performance in an independent dataset which is representative for the aimed application, that is, external validation. This is similar to assessing any type of test before using it in clinical practice.

From research clinical phenotyping and modeling to clinical applications

Beyond the scientific aim of making reliable datasets out of EHR for clinical research, AI is increasingly used for applications in clinical care. For instance, to predict development of disease or side effects, treatment response or to facilitate surgery and image interpretation.[34] As with any test or prediction, clinical application necessitates an even more rigorous model assessment.

Generalisability and implementation

The challenge of making even rigorously tested models that work well in clinical practice is exemplified by the epic sepsis model.[35 36] One of the most widely used clinical warning systems, the EPIC sepsis model was built on EHR data from 405 000 patient encounters across 3 health systems and was designed for use with real life EHR data. However, in a large external validation study, the Epic Sepsis Model failed to identify 67% (n=1709) of patients with sepsis.[37] Its failure in this independent testing is considered a result of lack of good external validation and a possible need for pragmatic clinical trials assessing the true impact.[38] Another reason might be that for implementation of EHR models, harmonisation of new EHR data, such as performed when combining dataset for science with the original datasets is not a standard procedure.

The EPIC sepsis model addresses the challenge of testing on one set of EHR data and applying it to a second. There are several reasons why a model can work across multiple institutions but not another: the data and population used to develop the model differs from the population where it is currently being applied. The EHR software itself can result in different codes for different conditions or laboratory studies. Differences in clinical practices between health institutions result in different types of noise, missingness and biases between EHR systems. The noisiness and missingness of data in EHR is not solely the result of different encounters with the hospital system due to different disease activity. Doctors' and patients' habits on frequency of visit request and additional examinations, insurance coverage, and the extent of information in medical notes result in strong batch effects between centres, doctors and perhaps patient groups. Accessibility of care and living distance to the clinic will influence the density of data in the EHR that could be correlated with patients' life circumstances and disease characteristics. These factors influence the performance of methods that were trained and tested on different EHR data. Traditional methods such as outlier detection to identify such problems are less suited for models that were built on high dimensional data. In addition, to test a model in a new system before implementation, it is advisable to monitor data shifts periodically and monitor impact in real time.[39 40] There are several guidelines for building and assessing AI both for algorithms building and assessment. Examples include Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis-AI, Standards for Reporting Diagnostic Accuracy (STARD)-AI and DECIDE-AI.[41–44] In addition, methods are being developed that allow a more automated approach for determining the equivalent codes across EHR systems to use in a model.[45]

In rheumatology, clinical research models exist for image interpretation, for example, erosion detection on X-rays, MRI interpretation, prediction of treatment failures and disease flares, picture-based synovitis detection but have not reached clinical implementation. Treatment response is particularly challenging, since both the documentation of disease activity, which is needed to define treatment response needs to be gleaned mainly from unstructured clinical notes, with a wide variation in how these concepts are documented.

Population health studies on EHR data

As outlined above, a big advantage of EHR data is that it could provide insights into disease aetiology and development. EHR data are often used for case series, nested case control studies and prospective and retrospective cohorts. Casey *et al* wrote an comprehensive overview of EHR studies that generated new insights into diseases, such as the association between chicken pox and stroke, neighbourhood deprivation and cardiovascular risk, and unconventional natural gas development and preterm birth.[46] In addition, biobanks linked with EHR data and samples for immuno and genotyping have further extended the reach and potential for translational research with RWD.[47 48]

Using EHR for population health studies does require special attention and caution to ensure high quality results. Differences in registration habits, disease severity, access to care and local healthcare standards influences the amount of noise, missingness and indication bias in the EHR. EHR data are in principle open cohorts, where people enter and leave at different moments during their disease course resulting in different density and lengths of trajectories. There are several reasons why EHR system can lack data on patients: patients have missed visits for personal or practical reasons, have been well or did not search for care, died, entered the system before digital registration was

available (leading to lack of baseline information (left-censoring)) or moved to a different the system (leading to right-censoring (lack of outcome information)). A valuable checklist to assess bias in population studies is the PROBAST tool.[49]

The length of the patients' trajectories influences the chance of being captured in case-control studies. When cases are randomly identified, for example, by using a certain drug at any time, the resulting dataset will be enriched with people who were doing well on those drugs and thus is biased towards good outcomes. To overcome this, a new user design or incident user design can be used, as is routinely performed in other types of RWD observational data, claims based research.[50] Here, patients are retrospectively selected at the time of drug prescription and all subsequent time points are part of the study. This temporal ordering protects studies against reversed causation.

Also the type of information that is registered for each patient is constrained by missingness. Clinicians' registrations are enriched with information that is useful for treatment and focuses on the interventions of the clinicians. Hereby, information including fundamental causes of diseases (social, environment, life-style) is less well registered.[51] These causes of missingness are systemic instead of at random, which can introduce bias if it is not taken into account. Simultaneously, the missingness or sparsity can be informative as well, for example, telling us a doctor was (not) suspecting a particular disease or a patient is (not) doing well. One study found that increased frequency of blood measurements, particularly during the late night early morning hours, had a strong correlation with mortality.[52] EHR is enriched for such associations, which can result in a reduction of analytical quality when ignored but could be an enrichment when used cleverly.[53] It does, however, require good domain knowledge and knowledge about the local healthcare system. This underlines the importance of involving clinicians into EHR studies.

Ethics

While the combination of ethics and legislation of EHR data usage is a subject on its own, we would like to address this topic in brief, as it is imperative before collecting and analysing any data. There are two main aspects that we would like to address, as they are directly pertinent to algorithms for phenotyping. As outlined, validation of any algorithm in the local EHR before broad clinical implementation is relevant to test the validity of the algorithms and the impact of possible error. For this, it is important to make the EHR data accessible for such analysis. Second, selection bias reduces generalisability of study results and the inclusiveness of EHR data offers a solution for biases in traditional designs. However, while EHRs often contain information on a broader population compared with recruitment-led studies, the evidence derived from EHRs will remain limited to the EHR population This on its own can be biased. This bias can relate to the way we obtain access to RWD. This necessitates a discussion on how to obtain data access in a manner where patients' rights are not violated and simultaneously we do not create additional research bias.

Legislation around data usage has reduced data accessibility. The current ideal (though not reality yet) is that the patient is the data owner and should provide access to their data.[54] Currently, General Data Protection Regulation (GDPR) requires a clear affirmative action in order to fulfil the consent criteria. This makes an automatic opt-in not possible (though it is not completely ruled out as an option). However, in addition to obtaining consent, there are several other situations where one is allowed to process personal data. These contain situations where one needs to fulfil a contract, there is a legal obligation, there is a vital interest, a public interest, in the exercise of official authority or when there is a legitimate (eg, commercial) interest provided it does not harm to the freedom and rights of the individual. Now it is allowed to subside the consent criteria for instance when it is not reasonably possible to obtain it (eg, when people died or the group of people is too large too reasonably be able to obtain the consent).

The problem with obtaining informed consent can be that it creates bias in patients who agree to consent (eg, by making the paperwork too difficult for certain groups).[55] Simultaneously if informed consent is required in any circumstance, we create a bias by excluding those who passed away. The questions whether this is ethical becomes even more pertinent when we are using RWD for developing algorithm for clinical practice.

Now it are not only clinicians or tech companies who realise the value of RWD. Also regulatory bodies are diving into the resource. RWD is increasingly as potentially powerful data to guide regulatory decision-making.[56] To do so requires the transformation of the noisy RWD to real world evidence (RWE). In recognition of the importance in developing the use of RWD to accelerate science, governments are helping to push the field forward by providing grants (EU, Horizon) aiming to accelerate science incorporating the use of RWD by setting out specific programmes and legislation.[1 57 58] The effectiveness of interventions can be studied in the complete variety of the true patient populations using EHR. This is one of the key reasons why FDA and the EU are focusing on exploring the validity of RWD: decision-making on the development, authorisation and supervision of medicines.[57] This generates an incentive for health authority to find solutions for the data access and consent problem. resulting into initiatives as the European Health Data space.[59] The opinions differ on whether this is an ideal and workable solution, which will also depend on the execution of the plan. At the least having an European wide solution and clarity on the interpretation of the law, would take away important current hurdles in the science with RWD.

CONCLUSION

In summary, EHRs provide a rich resource of RWD to advance our understanding of rheumatic conditions and when transformed to RWE can inform clinical care. EHR data complement traditional study designs because it captures almost the complete variety of patients, leading to more generalisable results. In addition, it is large, available and extensive in the type of data it captures. Using EHR data for science necessitates data cleaning and patient selection which requires different techniques than in observational cohorts or clinical trials, starting first by accurately classifying patients with the phenotype of interest. ML techniques provide high-throughput solutions for both patient phenotyping and to build prediction models. To ensure generalisability and prevent overfitting, validation in separate datasets and in each dataset over time is needed. As we move towards RWD-EHR data to guide clinical and regulatory decisions, academic–government–private partnerships are needed to determine the standards the data must meet, the ethics behind use of these data, and how the medical community will ensure that the algorithms remain relevant and continue to improve the health of the population they were developed to serve.

Contributors Both authors work on content and writing collaboratively.

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 applicable.

Provenance and peer review Commissioned; externally peer reviewed.

ORCID iD

Rachel Knevel http://orcid.org/0000-0002-7494-3023

REFERENCES

1 FDA. Available: https://www.fda.gov/media/120060/download
2 Adler-Milstein J, Holmgren AJ, Kralovec P, *et al*. Electronic health record adoption in US hospitals: the emergence of a digital "advanced use" divide. *J Am Med Inform Assoc* 2017;24:1142–8.
3 Ramirez FD, Motazedian P, Jung RG, *et al*. Sex bias is increasingly prevalent in preclinical cardiovascular research: implications for translational medicine and health equity for women: a systematic assessment of leading cardiovascular journals over a 10-year period. *Circulation* 2017;135:625–6.
4 The Lancet Rheumatology. The Lancet Rheumatology—tackling heterogeneity and embracing diversity. *Lancet Rheumatol* 2019;1:e1.
5 Benson K, Hartz AJ. A comparison of observational studies and randomized, controlled trials. *N Engl J Med* 2000;342:1878–86.
6 Coughlin SS. Recall bias in epidemiologic studies. *J Clin Epidemiol* 1990;43:87–91.
7 Hripcsak G, Albers DJ. High-fidelity phenotyping: richness and freedom from bias. *J Am Med Inform Assoc* 2018;25:289–94.
8 Bjornevik K, Cortese M, Healy BC, *et al*. Longitudinal analysis reveals high prevalence of Epstein-Barr virus associated with multiple sclerosis. *Science* 2022;375:296–301.
9 Choi MY, Hahn J, Malspeis S, *et al*. Association of a combination of healthy lifestyle behaviors with reduced risk of incident systemic lupus erythematosus. *Arthritis Rheumatol* 2022;74:274–83.
10 Chaganti S, Welty VF, Taylor W, *et al*. Discovering novel disease comorbidities using electronic medical records. *PLoS One* 2019;14:e0225495.
11 Arbuckle MR, McClain MT, Rubertone MV, *et al*. Development of autoantibodies before the clinical onset of systemic lupus erythematosus. *N Engl J Med* 2003;349:1526–33.
12 Reynolds KL, Arora S, Elayavilli RK, *et al*. Immune-related adverse events associated with immune checkpoint inhibitors: a call to action for collecting and sharing clinical trial and real-world data. *J Immunother Cancer* 2021;9:e002896.
13 Liao KP, Cai T, Gainer V, *et al*. Electronic medical records for discovery research in rheumatoid arthritis. *Arthritis Care Res* 2010;62:1120–7.
14 Carroll RJ, Thompson WK, Eyler AE, *et al*. Portability of an algorithm to identify rheumatoid arthritis in electronic health records. *J Am Med Inform Assoc* 2012;19:e162–9.
15 Bartels CM, Singh JA, Parperis K, *et al*. Validation of administrative codes for calcium pyrophosphate deposition: a Veterans administration study. *J Clin Rheumatol* 2015;21:189–92.
16 Hripcsak G, Albers DJ. Next-generation phenotyping of electronic health records. *J Am Med Inform Assoc* 2013;20:117–21.
17 Liao KP, Cai T, Savova GK, *et al*. Development of phenotype algorithms using electronic medical records and incorporating natural language processing. *BMJ* 2015;350:h1885.
18 Hsu J, Pacheco JA, Stevens WW, *et al*. Accuracy of phenotyping chronic rhinosinusitis in the electronic health record. *Am J Rhinol Allergy* 2014;28:140–4.
19 Maarseveen TD, Maurits MP, Niemantsverdriet E, *et al*. Handwork vs machine: a comparison of rheumatoid arthritis patient populations as identified from EHR free-text by diagnosis extraction through machine-learning or traditional criteria-based chart review. *Arthritis Res Ther* 2021;23:174.
20 Maarseveen TD, Meinderink T, Reinders MJT, *et al*. Machine learning electronic health record identification of patients with rheumatoid arthritis: algorithm pipeline development and validation study. *JMIR Med Inform* 2020;8:e23930.
21 GitHub. Available: https://github.com/Ievrex/DiagnosisExtraction_ML
22 PheKB. Available: https://phekb.org/
23 Sinnott JA, Cai F, Yu S, *et al*. PheProb: probabilistic phenotyping using diagnosis codes to improve power for genetic association studies. *J Am Med Inform Assoc* 2018;25:1359–65.
24 Liao KP, Sun J, Cai TA, *et al*. Tianxi CAI, with the million veteran program, high-throughput multimodal automated phenotyping (MAP) with application to PheWAS. *J Am Med Inform Assoc* 2019;26:1255–62.
25 Kashyap M, Seneviratne M, Banda JM, *et al*. Development and validation of phenotype classifiers across multiple sites in the observational health data sciences and informatics network. *J Am Med Inform Assoc* 2020;27:877–83.
26 Yu S, Ma Y, Gronsbell J, *et al*. Enabling phenotypic big data with PheNorm. *J Am Med Inform Assoc* 2018;25:54–60.
27 Van Calster B, Wynants L, Timmerman D, *et al*. Predictive analytics in health care: how can we know it works? *J Am Med Inform Assoc* 2019;26:1651–4.
28 Maurits MP, Korsunsky I, Raychaudhuri S, *et al*. A framework for employing longitudinally collected multicenter electronic health records to stratify heterogeneous patient populations on disease history. *J Am Med Inform Assoc* 2022;29:761–9.
29 Pournara E, Kormaksson M, Nash P, *et al*. Clinically relevant patient clusters identified by machine learning from the clinical development programme of secukinumab in psoriatic arthritis. *RMD Open* 2021;7:e001845.
30 Norgeot B, Glicksberg BS, Trupin L, *et al*. Assessment of a deep learning model based on electronic health record data to forecast clinical outcomes in patients with rheumatoid arthritis. *JAMA Netw Open* 2019;2:e190606.
31 Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* 1982;143:29–36.
32 He H, Garcia E. Learning from imbalanced data. *IEEE Trans Knowl Data Eng* 2009;21:1263–84.
33 Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS One* 2015;10:e0118432.
34 Bohr A, Memarzadeh K. The rise of artificial intelligence in healthcare applications. *Artif Intell Med* 2020:25–60.
35 Bennett T, Russell S, King J. Accuracy of the EPIC sepsis prediction model in a regional health system. Available: https://arxiv.org/abs/1902.07276
36 Gulshan V, Peng L, Coram M, *et al*. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA* 2016;316:2402–10.
37 Wong A, Otles E, Donnelly JP, *et al*. External validation of a widely implemented proprietary sepsis prediction model in hospitalized patients. *JAMA Intern Med* 2021;181:1065–70.
38 Habib AR, Lin AL, Grant RW. The EPIC sepsis model falls short-the importance of external validation. *JAMA Intern Med* 2021;181:1040–1.
39 Nalisnick E, Matsukawa A, Teh Y. Do Deep Generative Models Know What They Don't Know? In: *ICLR*, 2019.
40 Zadorozhny K, Thoral P, Elbers P. G CIN out-of-distribution Detection for medical applications: guidelines for practical evaluation. *arXiv* 2021.
41 Collins G, Dhiman P, Logullo P. TRIPOD-AI, 2021. Available: https://doi.org/10.17605/OSF.IO/ZYACB
42 Sounderajah V, Ashrafian H, Golub RM, *et al*. Developing a reporting guideline for artificial intelligence-centred diagnostic test accuracy studies: the STARD-AI protocol. *BMJ Open* 2021;11:e047709.
43 Vasey B, Nagendran M, Campbell B, *et al*. Reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. *Nat Med* 2022;28:924–33.
44 Norgeot B, Quer G, Beaulieu-Jones BK, *et al*. Minimum information about clinical artificial intelligence modeling: the MI-CLAIM checklist. *Nat Med* 2020;26:1320–4.
45 Hong C, Rush E, Liu M, *et al*. Clinical knowledge extraction via sparse embedding regression (KESER) with multi-center large scale electronic health record data. *NPJ Digit Med* 2021;4:151.
46 Casey JA, Schwartz BS, Stewart WF, *et al*. Using electronic health records for population health research: a review of methods and applications. *Annu Rev Public Health* 2016;37:61–81.
47 Bycroft C, Freeman C, Petkova D, *et al*. The UK Biobank resource with deep phenotyping and genomic data. *Nature* 2018;562:203–9.
48 Gaziano JM, Concato J, Brophy M, *et al*. Million veteran program: a mega-biobank to study genetic influences on health and disease. *J Clin Epidemiol* 2016;70:214–23.
49 Wolff RF, Moons KGM, Riley RD, *et al*. PROBAST: a tool to assess the risk of bias and applicability of prediction model studies. *Ann Intern Med* 2019;170:51–8.
50 Johnson ES, Bartman BA, Briesacher BA. The Incident User Design in Comparative Effectiveness Research. Effective Health Care Program Research Report No. 32. (Prepared under Contract No. HHSA290200500161). AHRQ Publication No. 11(12)-EHC054-EF. Rockville, MD Agency for Healthcare Research and Quality; 2012.
51 Link BG, Phelan J. Social conditions as fundamental causes of disease. *J Health Soc Behav* 1995;Spec No:80–94.
52 Agniel D, Kohane IS, Weber GM. Biases in electronic health record data due to processes within the healthcare system: retrospective observational study. *BMJ* 2018;361:k1479.
53 Gianfrancesco MA, Tamang S, Yazdany J, *et al*. Potential biases in machine learning algorithms using electronic health record data. *JAMA Intern Med* 2018;178:1544–7.
54 Mondschein CF, Monda C. The EU's General Data Protection Regulation (GDPR) in a Research Context. In: Kubben P, Dumontier M, Dekker A, eds. *Fundamentals of clinical data science*. Cham (CH): Springer, 2018.
55 Ohno-Machado L. Sharing data for the public good and protecting individual privacy: informatics solutions to combine different goals. *J Am Med Inform Assoc* 2013;20:1.
56 ASCPT. Available: http://ascpt.onlinelibrary.wiley.com/doi/full/10.1002/cpt.2479
57 CORDIS. Available: https://cordis.europa.eu/programme/id/H2020_SC1-DTH-12-2020
58 1st century cures act
59 EUR-Lex. Available: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52022PC0197