Digital rheumatology in 2022

https://doi.org/10.1038/s41584-022-00894-1

New developments in electronic health record analysis

Jutta G. Richter & Christian Thielscher

Electronic health records (EHRs) contain enormous amounts of real-world data that could inform researchers, doctors and patients about many aspects of rheumatology. However, EHRs are not yet fully utilized, mainly because automatic data extraction is difficult. Several studies in 2022 highlight the feasibility and clinical utility of computer-assisted EHR analysis.

The growth of digitalization — artificial intelligence, machine learning, big data, telemedicine and other new information and communication technologies (ICTs) — provides the potential to improve the diagnosis and treatment of patients with rheumatic diseases. Many ICTs are now entering clinical practice or are already part of standard care. For example, electronic health records (EHRs) and/or other patient documentation systems (such as in hospitals, practices and laboratories) offer a rich resource of data to advance our understanding of rheumatic conditions, and can complement traditional study designs because they capture almost the complete variety of patient journeys with real word data, leading to more generalizable results[1]. In addition, an increasing amount of data is being contained within these systems that might be used to analyse the epidemiological trends of inflammatory rheumatic diseases. However, difficulties remain in utilizing this data as EHR databases are typically partitioned into small entities, and extracting the data is challenging. Three studies in 2022 highlight promising approaches for addressing these issues, creating new epidemiological insights from big data and improving the feasibility and utility of EHR analysis[2–4].

Consolidation of EHR databases might help optimize EHR analysis to better capture epidemiology trends, as shown by Scott et al.[2]. To study the epidemiology of rheumatoid arthritis (RA), psoriatic arthritis (PsA) and axial spondyloarthritits (SpA) in England, the researchers analysed the Clinical Practice Research Datalink (CPRD) Aurum database, which contains longitudinal routinely collected EHRs from UK primary care practices. The database captures information ranging from demographic characteristics, diagnoses and symptoms, drug exposures to lab tests, and currently covers around 20% of the population in England, with a median follow-up time of ~9 years.

Scott and colleagues used algorithms and updated diagnostic codes, as well as synthetic DMARD code lists, to ascertain patients with a diagnosis of RA, PsA or axial SpA. This approach enabled the researchers to calculate the annual incidence and point-prevalence of RA, PsA and axial SpA diagnoses from 2004 to 2020, stratified by age and sex. For example, the point-prevalence of RA and PsA diagnoses increased annually from 2004 onwards, peaking in 2019, before falling slightly. The point-prevalence of axial SpA diagnoses increased annually (except in 2018 and 2019), peaking in 2020. Finally, the annual incidence of RA, PsA and axial SpA diagnoses fell by 40.1%, 67.4%, and 38.1%, respectively between 2019 and 2020, probably reflecting the impact of the COVID-19 pandemic. This type of insight is especially useful for planning and shaping health services (in this case, NHS services) particularly for the elderly population. Similar approaches could be used in other health-care systems to plan accordingly.

In many situations, automatically extracting data on patients with a certain diagnosis from a database and/or defining subgroups of patients using this data can be useful for researchers. Zheng et al.[3] studied the ability of the Phenotype KnowledgeBase (PheKB) algorithm to automatically identify patients with RA from an EHR database. They found that the specificity of this algorithm was quite good (95%), but the sensitivity was poor (~72%). Notably, the sensitivity of this algorithm was especially low in patients with seronegative RA. The phenotyping algorithm used an automated calculation (based on penalized logistic regression) to select clinically relevant features. Various useful features were captured by the algorithm (such as International Classification of Diseases (ICD) codes and rheumatoid factor laboratory test results), but others were missed, including anti-citrullinated protein antibody (ACPA) laboratory test results and text-based indications of joint involvement. In addition, the phenotyping algorithms were unusable for a notable number of patients owing to a lack of data in the necessary structured format. Hence, the results indicate that ability of this platform to identify the key data elements needed to define phenotypes is limited and expert input is still required. These findings highlight the need for careful design choices when developing phenotyping

Key advances

- Consolidation of electronic health record (EHR) databases can improve our understanding of real-world patient journeys[2].
- A newly developed natural language processing (NLP) pipeline can automatically extract outcome measures from EHR databases and has potential for EHR data analysis for both clinical and research purposes[4].
- Although an automated phenotyping algorithm has potential for diagnosing patients with certain subtypes of rheumatoid arthritis using EHR data, design choices (such as the definition of what key elements are used for phenotyping) and missing data currently limit its clinical utility[3].

algorithms. Before phenotyping algorithms can be implemented in routine care, approaches for handling missing data are needed.

Although most EHR systems include some structured data fields for capturing particular information (such as ICD codes), the included fields and their usability can vary across systems, the majority of EHR data are often documented in an unstructured format (such as text) and are thus difficult to analyse. The study by Humbert-Droz et al.[4] highlights one method for navigating this issue. Using data from 2015–2018, including 34 million notes from 854,628 patients, 158 practices and 24 EHRs, the researchers developed and evaluated a natural language processing (NLP) pipeline for extracting mentions of rheumatoid arthritis outcome measures and scores from free-text outpatient rheumatology notes within the Rheumatology Informatics System for Effectiveness (RISE) registry. The RISE registry combines data from different EHRs and consolidates them. The NLP pipeline had a good internal and external validity, with a sensitivity, positive predictive value and F1 score of 95%, 87% and 91%, respectively. Substantial agreement was observed between the scores extracted from the RISE notes and scores derived from structured data within the RISE registry. Thus, the pipeline has potential for facilitating outcome measurement in research but also in clinical care. In the future, the NLP pipeline might also support personalized medicine if used, for example, to automatically analyse the historical EHR data of a specific patient.

Rheumatological diseases are typically chronic in nature. Over time, EHRs can gather enormous amounts of data on individual patients that are difficult to track for human doctors but might provide very helpful information. Putting together EHR data in ever-increasing databases helps to improve research (for example, epidemiology research), and tools such as the NLP pipeline should enable automatic access to this rich resource. In summary, the use of artificial intelligence and machine learning algorithms will hopefully lead to optimized patient-centred care in the near future.

Jutta G. Richter[1] & Christian Thielscher[2]

[1]Department for Rheumatology and Hiller Research Center, University Hospital, Medical Faculty of Heinrich-Heine-University Duesseldorf, Duesseldorf, Germany. [2]Competence Center for Medical Economics, FOM University, Essen, Germany.

e-mail: christian.thielscher@fom.de

Published online: 19 December 2022

References

1. Knevel, R. & Liao, K. P. From real-world electronic health record data to real-world results using artificial intelligence. *Ann. Rheum. Dis.* https://doi.org/10.1136/ard-2022-222626 (2022).
2. Scott, I. C. et al. Rheumatoid arthritis, psoriatic arthritis, and axial spondyloarthritis epidemiology in England from 2004 to 2020: An observational study using primary care electronic health record data. *Lancet Reg. Health Eur.* **23**, 100519 (2022).
3. Zheng, H. W. et al. Evaluation of an automated phenotyping algorithm for rheumatoid arthritis. *J Biomed Inform.* **135**, 104214 (2022).
4. Humbert-Droz, M. et al. Development of a Natural Language Processing System for Extracting Rheumatoid Arthritis Outcomes From Clinical Notes Using the National Rheumatology Informatics System for Effectiveness Registry. *Arthritis Care Res. (Hoboken)* https://doi.org/10.1002/acr.24869 (2022).

Competing interests

The author declares no competing interests.

Related links

The Clinical Practice Research Datalink (CPRD) Aurum database: https://cprd.com/cprd-aurum-march-2021

The Rheumatology Informatics System for Effectiveness (RISE) registry: https://www.rheumatology.org/Practice-Quality/RISE-Registry