Operational challenges of building a million-patient cohort from EHRs: The COhort of DIabetic patients (CODIA) on the AP-HP EDS

Judith Abécassis, Théo Jolivet, Audrey Bergès, Elise Liu, Jean-Baptiste Julla, Yawa Abouleka, Julie Alberge, Isabel Bonnetier, Thomas Petit-Jean, Romain Bey, et al.

Operational challenges of building a million-patient cohort from EHRs: The COhort of DIabetic patients (CODIA) on the AP-HP EDS

Judith Abécassis[1], Théo Jolivet[2], Audrey Bergès[2], Elise Liu[2], Jean-Baptiste Julla[2], Yawa Abouleka[2], Julie Alberge[1], Isabel Bonnetier[2], Thomas Petit-Jean[2], Romain Bey[2], Candice Estellat[2], Florence Tubach[2], Gaël Varoquaux[1], Louis Potier[2]

[1] Inria Saclay-Ile-de-France, Soda team, Palaiseau, France
[2] Assistance publique-Hôpitaux de Paris, Paris, France

Abstract

Electronic health records (EHRs) are becoming increasingly accessible as hospitals set up health data warehouses. Those data come with many promises for applications in clinical, epidemiologic, and translational research, but also with particularities in the data access, the data collection process, and the data format that must be accounted for in the preparation and the subsequent analyses. This brief article shares our practical experience building a solid and exhaustive cohort of diabetic patients in the Greater Paris University Hospitals Clinical Data Warehouse (AP-HP EDS). This ambitious project's substantial engineering and computing resources should be balanced with an adequate sharing and reusing policy for future studies.

Introduction

The Greater Paris University Hospitals (AP-HP for Assistance Publique-Hôpitaux de Paris) have made routine care data available for research purposes within its Clinical Data Warehouse (CDW, Entrepôt de données de santé EDS). The EDS consists of a varied collection of structured and unstructured data, including demographic data, biology results, imaging exams, medical notes, drug prescriptions, and administrative encoding of the patient path collected for billing purposes. Before it is made accessible to researchers, the data is pseudonymized, systematically altering names or identifying information for all data types [1]. The EDS holds information about 11 million patients spanning 38 hospitals.

In the CODIA (**Co**hort of **Dia**betic patients) project, we propose to build a large longitudinal cohort of patients with diabetes, a progressive disease that affects around 10% of adults worldwide and leads to severe complications (blindness, kidney failure, heart attacks, stroke, and lower limb amputation) and premature death. The objective is first to describe and predict the possible evolution trajectories of patients to understand potential risk factors and then try to assess the effect of medical interventions on those trajectories, depending on patients' characteristics. Beyond those planned analyses, CODIA intends to become a comprehensive resource for further studies on diabetes.

Legal procedure to access the data

The authorization of the creation of the AP-HP EDS by the CNIL in 2017 (Commission nationale de l'informatique et des libertés, the French administration in charge of protecting data privacy and individual liberties) has allowed a simplified procedure to access the data through the AP-HP CSE (Comité Scientifique et Ethique, scientific and ethical board) if you are a healthcare provider at the AP-HP. The CSE meets once a month, and overall, the estimated delay to access the data is around 3

months. This delay can be lengthened if several institutions are involved, which is our case with AP-HP and Inria, but is simplified by a more general partnership, the Bernoulli Lab, with a pre-established master agreement.

The initial request must describe and substantiate the requested data and the level of anonymity. In particular, dates are usually blurred to prevent the further re-identification of patients. The patients are allowed to opt out of using their data in the EDS.

Practical implications of cohort selection

The cohort definition step is critical to the validity of any result obtained from the cohort as it conditions the definition of the population of interest for subsequent studies [2]. Indeed, systematically missing patients can bias the analyses conducted with the cohort, and including too many patients can significantly improve the burden on limited computing resources and hinder the administrative authorizations to access the data, as limiting the size of datasets is one of the actionable levers to ensure patients' privacy.

The EDS has implemented a tool for cohort selection called Cohort360 to explore the choice of patients along various criteria. The recommended way to access Cohort360 is to collaborate with data scientists in URCs (Clinical Research Units). Still, over time, we've appreciated having direct access to better control of the inclusion criteria. The diabetic status (and the diabetes subtype) is not recorded in a consolidated and consistent way throughout the EDS and must be inferred from other available data representing the patient phenotyping step. The most direct information is to use the CIM-10 diagnosis codes used for billing in the case of inpatients. Hence, codes are missing for outpatients and deemed unreliable [3]. In practice, we observe that depending on the visit’s primary reason, diabetes is not always present in the codes, and there are often errors in the diabetes subtype code [4].

At the cohort extraction step, the data is -by construction- not yet accessible, so complex approaches involving machine learning classification algorithms cannot be applied, and selection criteria must be based on existing variables. However, Cohort360 only implements a subset of the possible requests. As a result, our initial cohort consisted of two cohorts we had to merge during our analyses. One of our criteria was to include patients who had been to a diabetes or endocrinology department. Still, on closer inspection, we realized that we could not select departments, only larger care units, which led to the inclusion of many non-diabetic patients. Also, the list of relevant care units was challenging to maintain, and so many services were missed. With this knowledge, we modified the cohort definition, relying instead on regular expressions to find evidence of a diabetes diagnosis or treatment mentioned in the medical records, in addition to criteria based on biological measures of glycemia or the administration of anti-diabetic treatments, and ICD-10 diagnosis codes present only in inpatients.

Data quality considerations

The fact that research with the EDS data is a secondary utilization has significant consequences: first, the collected data is guided by the patient needs, the healthcare practitioner's specialty (e.g. facing the same patient, a geriatric doctor will report key elements regarding the autonomy of their patient,

while a cardiologist will focus on a cardiac examination), the local habits for diagnosis coding and reporting [3], the status of the patient (inpatient with medication and biology measurements reported in a structured tabular format or outpatient for which only medical notes are introduced in the system). On top of those inevitable sources of bias, the EDS is intrinsically linked to the hospital information system; hence, the temporal depth of data available depends on the adoption date of a new hospital software that varies for each service of the AP-HP hospitals [5]. Also, while the EDS was created in 2017, the structured biology measurements are only available as of 2020. Finally, some critical information is collected from external sources, like the vital status of the patients, obtained from the CépiDc database, maintained by the Insee (the National Institute of Statistics and Economic Studies), but is integrated with a substantial time lag, resulting in unreliable death status over the cohort. At the same time, mortality is frequently used as a target variable [6]. Moreover, by construction, the data only covers visits and events at an AP-HP hospital [4, 6].

Going further and enriching data using NLP

The CODIA cohort contains 1,264,434 patients for 14,633,336 visits. Nonetheless, behind those very large numbers, there is a substantial disparity in the available data between patients, depending on the frequency and the nature of hospital visits (Table 1). Around half of the recorded visits are outpatient consultations, for which almost no structured data is collected. This current phenomenon in analyzing EHRs can be partly overcome by extracting relevant information from medical notes or prescriptions [7]. We have applied a deep learning language model fine-tuned on the AP-HP data to perform Named Entity Recognition (NER) to identify drug names, biology, and vital sign measures. Our preliminary results suggest that this procedure allows us to collect biology information for more than 80% of the patients instead of 57% if we rely solely on structured data.

This step has been challenging, as it requires substantial skills to run the model (developed within the data science team of the AP-HP EDS) and adapt it to the particular computer infrastructure of the project and the volume of more than 75 million medical notes of our cohort. The computing resources in the secured environment of the EDS are limited, resulting in a total time of 3 to 4 weeks to apply this NER model to the entire dataset with 2 T4 GPUs. The next step is to use another model to extract and characterize the medical conditions of the patients in the cohort [8] and to perform entity linking of the found drugs and biology measures to standardized nomenclatures (for instance, ATC codes for drugs). This is the entity linking step, which requires much richer data annotation.

	structured data		NLP-derived data	
	count	unique patients	count	unique patients
Drug prescriptions	46 158 851	725 836	216 151 894	1 095 729
Drugs administered in the hospital	159 831 263	485 713		
Biology results	694 402 621	723 469	288 310 457	1 046 643
Medical notes (text)	76 869 510	1 247 333		
Visits	14 633 336	1 264 434		
Consultations	7 001 404	1 153 908		

Table 1: Main characteristics of the available information in the CODIA cohort (preliminary results)

Challenges and future directions

Our experience building a database of EHRs from diabetic patients in the AP-HP EDS illustrates this data's fantastic potential to conduct unprecedented observational analyses. However, this enthusiastic perspective of data reutilization should come with a realistic evaluation of the required effort: several months to access the data (and more if a partnership must be concluded among several institutions), the need to have solid expertise to select the patients adequately, which might require several attempts, and to use natural language processing models to extract features from the medical notes in a GPU-poor environment. This effort can be difficult to achieve for each single study, explaining our choice to build an exhaustive and robust resource that will be reused over multiple analyses. This additional effort involves better coding practices to enhance the robustness of the code of the project [9]. It is important to think of better ways to mutualize and share this preparation effort, to promote the re-use of this data, while maintaining the adequate level of privacy, as is done in other fields [10].

Acknowledgments

JA, TJ, JA, and GV were partially supported by the European INTERCEPT-T2D project.

References

[1] Tannier, X., Wajsbürt, P., Calliger, A., Dura, B., Mouchet, A., Hilka, M., & Bey, R. (2024). Development and validation of a natural language processing algorithm to pseudonymize documents in the context of a clinical data warehouse. Methods of Information in Medicine.

[2] Richardson, W. S., Wilson, M. C., Nishikawa, J., & Hayward, R. S. (1995). The well-built clinical question: a key to evidence-based decisions. ACP journal club, 123(3), A12-3.

[3] Chawla, N., Yabroff, K. R., Mariotto, A., McNeel, T. S., Schrag, D., & Warren, J. L. (2014). Limited validity of diagnosis codes in Medicare claims for identifying cancer metastases and inferring stage. Annals of epidemiology, 24(9), 666-672.

[4] Haneuse, S., & Daniels, M. (2016). A general framework for considering selection bias in EHR-based studies: what data are observed and why?. EGEMs, 4(1).

[5] Remaki, A., Playe, B., Bernard, P., Vittoz, S., Doutreligne, M., Chatelier, G., ... & Bey, R. (2023). Adjusting for the progressive digitization of health records: working examples on a multi-hospital clinical data warehouse. medRxiv, 2023-08.

[6] Beesley, L. J., & Mukherjee, B. (2022). Statistical inference for association studies using electronic health records: handling both selection bias and outcome misclassification. Biometrics, 78(1), 214-226.

[7] Khurshid, S., Reeder, C., Harrington, L. X., Singh, P., Sarma, G., Friedman, S. F., ... & Lubitz, S. A. (2022). Cohort design and natural language processing to reduce bias in electronic health records research. Npj Digital Medicine, 5(1), 47.

[8] Petit-Jean, T., Gérardin, C., Berthelot, E., Chatellier, G., Frank, M., Tannier, X., ... & Bey, R. (2024). Collaborative and privacy-enhancing workflows on a clinical data warehouse: an example developing natural language processing pipelines to detect medical conditions. Journal of the American Medical Informatics Association, 31(6), 1280-1290.

[9] Williams, R., Kontopantelis, E., Buchan, I., & Peek, N. (2017). Clinical code set engineering for reusing EHR data for research: a review. Journal of Biomedical Informatics, 70, 1-13.

[10] Girard-Chanudet, C. (2023). La justice algorithmique en chantier. Sociologie du travail et des infrastructures de l'Intelligence Artificielle (Doctoral dissertation, EHESS (École des Hautes Études en Sciences Sociales)).