Electronic health records for the diagnosis of rare diseases

Nicolas Garcelon, Anita Burgun, Rémi Salomon, Antoine Neuraz

Electronic health records for the diagnosis of rare diseases

Nicolas Garcelon[1,2], Anita Burgun[2,3], Rémi Salomon[1,4], Antoine Neuraz[2,3]

[1] INSERM, UMR 1163, institut Imagine, université Paris-Descartes, Sorbonne Paris-Cité, 75015 Paris, France
[2] INSERM, Centre de Recherche des Cordeliers, UMRS 1138, eq 22, Université Paris Descartes;
[3] Hôpital Necker - Enfants Malades, Department of Medical Informatics, Assistance Publique - Hôpitaux de Paris (AP-HP), France.
[4] Hôpital Necker - Enfants Malades, Department of Pediatric Nephrology, Assistance Publique - Hôpitaux de Paris (AP-HP), France.

Abstract:

With the emergence of electronic health records, the reuse of clinical data offers new perspectives for the diagnosis and management of patients with rare diseases. However, there are many locks for the repurposing of clinical data. The development of decision support systems depends on the ability: to recruit patients; then to extract and integrate the patients' data; to mine and stratify these data; and to integrate the decision support algorithm into the patient care. This last step requires an adaptability of the electronic health records to integrate learning health system tools. In this literature review we examine the research that provide solutions to unlock these barriers and accelerate translational research: structured electronic health records and free text search engine to find patients, data warehouses and natural language processing to extract the phenotypes, machine learning algorithms to classify patients, similarity metrics to diagnose patients etc. Medical informatics is experiencing an impellent request to develop decision support systems and this requires ethical considerations for clinicians and patients to ensure a good usage of health data.

Keywords: pediatric nephrology, education

Introduction:

Since the rise of the electronic health record (EHR), clinicians spend an increasing number of hours at their computers, from 49% of their total time[1] up to 66%[2]. Moreover, hospital adoption of EHRs with advanced functionality has increased in the last decade[3]. This is why "in the current digital age, the electronic health record represents a massive repository of electronic data points representing a diverse array of clinical information"[4]. Such data repositories can be used for research purposes, to accelerate the recruitment of patients, and discover new knowledge. Moreover, being able to leverage this massive amount of latent knowledge to support doctors' decision making is essential.

In 2011, Frankovitch et al. reported on how they used the clinical data warehouse of Stanford hospital to make a clinical decision for a young patient with systemic lupus erythematosus complicated by nephrotic-range proteinuria, antiphospholipid antibodies, and pancreatitis[5]. There were neither guidelines nor publications which could help to make this decision. They queried the data warehouse for similar cases and made "the decision on the basis of the best data available." This paper opened the gates to a new utilization of EHRs: repurposing the clinical data to develop decision support systems inside the EHR. The virtuous circle[6] of translational research could then be closed: from routine care data to research, and from research to clinical care.
In the context of rare diseases, the capacity of reusing EHR data to support diagnosis is even more critical. Indeed, given the rarity of these diseases, the diagnosis journey of the patients often last for years[7]. Tools taking advantage of the data of already diagnosed patients to help the decisions for the new cases could save a lot of time.

We designed a query (Figure 1) to find references in Medline on how the EHR may be used to diagnose rare diseases. This query returned 52 references as of February 14 2019. We performed a manual review and removed 7 references that were false positives (acronyms with other meaning). Among the remaining 45 publications (categories are not exclusive):

- 20 publications report the results of clinical studies (epidemiological studies, clinical trials, natural history of diseases). The authors used EHRs to recruit the patients and/or to describe the natural history of the diseases.
- 5 publications present structured EHRs tailored for the reuse of the data for a specific rare disease.
- 6 publications describe rare disease registries.
- 5 publications provide description of methods to recruit patients based on EHR data.
- 4 publications explain how they used EHRs to extract gene/phenotype association or disease/environment association.
- 1 publication provides a global overview of the repurposing of the EHR.
- 3 publications correspond to online applications for the diagnosis of rare diseases.
- Only 4 publications describe methods to diagnose rare diseases by repurposing the EHR data, but none of them are integrated in any EHR system.

The paucity of publications found may be explained by several issues, and at least 6 steps must be performed for developing a clinical decision support system (CDSS) and integrating it into the EHR. These steps include (Figure 2):

1. The identification and recruitment of patients (which may require manual annotation/review of the patient records),
2. The extraction of information from heterogeneous clinical data (e.g. narrative reports, imaging system, lab exams),
3. Data integration, enrichment and data auditing to assess data quality,
4. Development of machine learning methods,
5. The evaluation of the CDSS,

6. The integration of the CDSS in the EHR system, and its usability, so that it can be used by the doctor during patient care process.

In addition, Shortliffe et al. highlighted six characteristics a decision support system must have in the clinical field: understandability, time efficiency, ergonomics, adaptability for the clinical domain, respect for users, and solid scientific reliability[8].

Similarly to the publication by Jensen et al. in 2012[9], we expanded our review on approaches proposed to address these issues, but we prioritized the publications related to rare diseases or genetic diseases. Concretely we selected publications related to the recruitment of patients through the EHR: (EHR or electronic health record) and ("cohort identification" or "cohort selection" or "patient identification" or "recruitment").We then focused on the technologies used to improve this selection by adding “data warehouse,” then “natural language processing", "Knowledge extraction", "deep phenotyping", "rare disease and diagnosis", "learning health system" and "Genomics".

Identification and recruitment of patients

Several approaches to help identifying patients for clinical studies have been explored. Most of the time investigators used legacy terminologies like the International Classification of Diseases (ICD) to identify the patients of interest[10]. This method is barely applicable for rare diseases. Fung et al. estimated the coverage of the 6,500 rare diseases up to 11% with ICD-9-CM and 21% with ICD-10[11]. EURORDIS was even more pessimistic and estimated that only 500 rare diseases have a specific code in the ICD10[12]. Since the coverage of ICD codes for rare diseases is very low, it is necessary to explore other data sources.

EHR for recruitment and clinical research

The repurposing of EHRs performed better for patient recruitment than using data from registries[13]. Schreiweis et al. compared the abilities of five EHRs to recruit patients in clinical trials[14]. They did not find any functional components to support the research requirement of patient recruitment. For this reason, several authors developed dedicated EHRs tailored for specific rare diseases and data repurposing for research. Vawdrey et al. proposed a model based on Common Data Elements to structure the EHR for supporting clinical research[15]. The OpenEHR initiative integrated genomics archetypes in their EHR model and showed their applicability to clinical practice with rare genetic diseases[16]. Other solutions propose to map the EHR’s local logical schema onto a common external schema to facilitate data mining and extraction[17].

This last example introduces the necessity of integrating data into a new data model to clean and enhance it for efficient repurposing. EHR data models are designed for care purposes not for research. They are optimized to efficiently enter and retrieve all data for one patient at a time. Therefore, they are composed of hundreds to thousands of different tables which make it

difficult to make transversal queries (i.e. query specific data for a large number of patients). The strategies that were implemented to overcome this issue: (i) the creation of structured EHRs dedicated to translational research and (ii) the creation of copies of the clinical production databases in translational data warehouses, are described below.

Structured EHRs for translational research

The first strategy developed to enhance the repurposing of clinical data produced during patient care was to adapt the EHR for this objective.
One possible way is to develop specialized EHRs for specific disciplines or diseases. Bremond-Gignac et al. showed the benefit of an ophthalmology-specific EMR for collecting a comprehensive ocular visual phenotype useful for clinical research[18]. Other approaches consisted in adding disease oriented predefined templates to existing EHRs, for example "infantile spasms template"[19].
On the other hand, a horizontal approach would consist in finding a way to explore all the diseases, including rare diseases, in a common format. The French consortium for rare diseases (BNDMR) defined a minimum data set common to all rare diseases[20]. They used international terminologies such as the Human Phenotype Ontology (HPO)[21], the Online Mendelian Inheritance in Man (OMIM)[22] and Orphanet[23], as well as a standard data exchange format, the Fast Healthcare Interoperability Resources (FHIR) to structure and transfer the data[24]. Semantic integration of EHRs and other resources is another way to improve data reuse. Chelsom et al. enhanced the open source EHR cityEHR with semantic web technologies OWL (Web Ontology Language) to map the data onto knowledge databases in order to display relevant information to the clinician[25]. OSSE[26] or FluxMed[27] also implement structured EHRs specialized for rare diseases.
Asking clinicians to tick hundred of boxes is, however, not efficient: it is time consuming and in the context of rare diseases checkboxes do not qualify diagnoses or phenotypic descriptions. We will detail the importance of free text later in the article. Nevertheless, if a structured specialized-EHR is required for a rare disease, it is no longer acceptable not to use an international thesaurus like HPO or SNOMED-CT[28]. Working on several databases is not an issue if they are able to interoperate through a common thesaurus.

Translational data warehouses

The common EHR system is rarely the sole source of information in a hospital. Most of the time the Hospital Information System consists of a combination of several software and databases dedicated to specific tasks or specialties (EHR, PACS, Biological Results, pathology, etc.). These databases are more or less interoperable. In addition, dozens of decommissioned databases are stored on local hard drives or servers inside the hospital and are generally not available for clinical care or research. Until recently, physicians and researchers have built registries to serve specific goals. Wen et al. compared the efficiency (time, cost and number of participants) of registry versus data warehouse methods to recruit patients. They demonstrated that the data warehouse method was more efficient with more participants recruited in less time and at a lower cost[29]. Built on top of EHRs, Clinical Data Warehouses enable collection and secondary use of clinical data for many purposes, including

research (Phenome-wide analysis, record mining, epidemiological surveillance, pharmacovigilance, etc.).

Most major hospitals built their own data warehouses to aggregate all the clinical data produced[30–33]. For example, the multidisciplinary consortium NACI (NephCure Accelerating Cures Institute) developed a data warehouse specifically to improve care for nephrotic syndrome[34].

Several research projects have developed open source or free license data warehouses. The most widespread and commonly used is i2b2 developed by the NIH-funded National Center for Biomedical Computing based at Partners HealthCare System in the US[35,36]. i2b2 is primarily oriented toward structured data related to thesauri. Another open source tool, Dr Warehouse was developed at Necker hospital, a pediatric hospital specializing in rare diseases, and was oriented toward clinical documents such as narrative reports[37]. It integrates 26 different sources of clinical data, over 500 000 patients and 5 millions clinical reports. These data warehouses aim at accelerating recruitment but also the mining of data by integrating pre- and post-processing methods. For example, Davis et al. developed natural language processing methods to extract clinical traits from the Vanderbilt data warehouse in order to recruit patients with multiple sclerosis[38]. Their algorithm was highly specific to cases of multiple sclerosis but it highlights the potentiality of biomedical data warehouses to recruit patients with complex phenotypes.

OHDSI[39] (formerly the Observational Medical Outcomes Partnership - OMOP) initiative contributes to standardize data warehouses by building and maintaining the OMOP CDM (Common Data Model)[40]. This standard would allow interoperability and reproducibility of studies on different data warehouses.

Natural Language Processing methods

Several studies showed that narrative reports contain more information than coded data[41–43]. Especially for rare diseases or undiagnosed patients, free text is the only way to describe the fine-grained signs and symptoms of the patients without nosological guidelines. An important element of the text may reference the patient's family history or note the absence of clinical signs. Indeed, we showed that a query on rare disease records may return 70% false positive results due to sentences contained negated terms (e.g. "the patient does not have renal insufficiency") and family history (e.g. "The father has Crohn's disease") [44]. Several methods were developed to detect negation, experiencer, and temporality in clinical narrative, including Negex, Context, DEEPEN, and NegAIT[44–49]. Most of these methods are based on regular expressions, thus not requiring a machine-learning phase on an annotated corpus of texts. They show high levels of accuracy, recall and precision. One main objective of natural language processing methods is to reduce the noise of the search engine results and to make pre-screening work easier and less time consuming[50]. A second purpose is to structure free text to facilitate data mining and knowledge extraction. It will be presented in the "Knowledge discovery on rare diseases" section.

Search engine on EHR data

The first use case of biomedical data warehouses is to find patients eligible for biomedical studies (feasibility study, pre-screening) or multidisciplinary consultation meetings (patients like mine). The amount of time required to find patients without a suitable tool is untenable, and even impossible for some criteria combining multiple sources of clinical data (e.g. creatininemia > 2 mg/dl AND Rituximab AND Uveitis AND Cataract). The strength of data warehouses is to make this type of task possible and fast by integrating all of the data available within the hospital. In Nephrology, Schmidt et al. developed a method based on faceted searches allowing a user to narrow down searches on items that co-occurred in the same document[51,52]. i2b2 provides a search engine through a thesaurus allowing the user to build complex queries on structured data[36]. Other search engines, like in Dr Warehouse[37] and Emerse[53], are oriented toward free text and provide an intuitive user interface to find patients. Dr Warehouse combines queries on free text, coded data, and temporal relations between the search criteria.
PheKB is a platform that proposed a catalog of queries including criteria to find patients based on phenotypes (combination of ICD codes, medications, laboratories and natural language processing)[10]. For example, Mayo clinic share a flowchart to recruit patient called "Electronic Health Record-based Phenotyping Algorithm for Familial Hypercholesterolemia".

Knowledge extraction from EHR

Our second category of use cases regards the extraction of information and/or knowledge from EHRs. With focus on rare diseases, two specific tasks will be explored: the extraction of phenotypes and their associations, and the analysis of clinical pathways.

Phenotype extraction and association test

Kohane introduced the EHR-Driven Genomic Research (EDGR) as the repurposing of the EHR for genomic characterization[54]. Once clinicians are able to create a cohort of patients, the next challenge is to mine the available data. The difficulty lies in extracting information from free text. Indeed, narrative clinical reports hide relevant information, which can be automatically detected by machine learning algorithms within NLP pipelines[55–60]. The extracted medical concepts are then mapped to the Unified Medical Language System (UMLS) Metathesaurus® developed by the US National Library of Medicine.
This concept extraction pipeline can be used to generate high throughput phenotyping on a cohort of patients. Comparing the result to the data available in the literature (Orphanet, Wikipedia, PubMed), researchers can identify novel associations and early phenotypes[13,43,56,61–63]. Pairing with control, phenome-wide association studies (PheWAS) can be automatically processed to test association between SNPs and extracted phenotypes[64,65]. Photographs of patients can also be used to identify phenotypes by using deep learning[66]. The application of artificial intelligence to the analysis of patients' images can facilitate the discovery of new genetic disorders[67].

Clinical Pathways analysis

Another dimension to explore is the sequence of the interactions between the patient and the hospital or the health system in general, also known as clinical pathway. Clinical pathways of the patients inside the hospital may also be of interest to describe a cohort of patients[68–70]. In the context of chronic kidney disease (CKD), Zhang and Patman identified seven subgroups of patients by using data-driven clinical pathway learning[71] and developed a prediction algorithm to predict the most probable clinical pathway given a patient's characteristics.

Diagnosis support system tools

Machine Learning

Machine learning can be used in each step of the pipeline of creation of a CDSS: to recruit the patients and to extract knowledge from the narrative records, and to classify the patient with the right diagnosis.

Machine learning algorithm can be divided into 4 categories :

- Supervised learning algorithm: It requires labeled data for the learning process. It produces a predictive model for numeric prediction (linear regression, KNN, Gradient Boosting) or classification (logistic regression, random forect, Support Vector Machine, decision tree) or both (neural network).
- Unsupervised learning algorithm: it does not require labeled data. It produces descriptive model. It is used for Phenome/Genome Wide Association (Association rule learning algorithms), clustering (K-Means, Vector Space Model), or to learn a representation for dimensionality reduction (autoencoder neural network).
- Semi-supervised learning algorithm: it combines labeled and unlabeled data. It produces predictive (classification) and descriptive model (clustering).
- Reinforcement learning algorithm: it produces predictive model (classification) and control. It is first trained with the training data to complete a specific task. Then, it continuously learns in interaction with the environnement to optimize the reward and minimize the risk to complete this task. For example, Markov Decision Process is a reinforcement learning. But reinforcement learning is data hungry, this creates problems for applying this model in rare diseases.

Neverthless, machine learning algorithms require a large quantity of data to learn the expected outcome. In the context of rare diseases, this number is generally not reachable. The best approach is to combined rule-based method and machine learning algorithms.

Internet tools and diagnosis

The web has become a primary source of information about rare diseases. Several websites offer search engines specializing in rare diseases and use scientific literature to allow users to find diagnoses by querying phenotypes (*FindZebra*[72], *Phenolyzer*[73], *PhenoDis*[74], *Rare disease discovery*[75], *GDDP*[76]). They extract and curate information found in publicly available

information on rare diseases such PubMed, Wikipedia or Orphanet. And they enriched them with ontologies.
HPO is widely used as a standard ontology to annotate diseases with phenotypes, to calculate the information content of phenotypes, or to provide common pivot semantics to heterogeneous databases. Other knowledge resources are used to enrich existing data, for example rare disease description or gene-phenotype annotations: Orphanet[23], UMLS[77], eRAM[78], Uberpheno ontology[79], Gene Ontology[80], ClinVar[81] etc. Machine learning methods integrate those databases in the learning phase to compute similarity distances, phenotype scores, and to provide a list of prioritization of causal variants (xRare[82], RDAD[83] and phenotype risk score[84]).

EHR and diagnosis

The CDSS integration may either rely on the EHR structured data alone or exploit the whole spectrum of EHR data. Wulf et al integrated, in openEHR, a rule based algorithm using clinical structured data within the EHR to early detect systemic inflammatory response syndrome[85].
Text mining is needed to exploit narrative clinical data [86]. EHR-*Phenolyzer*, developed by Son et al., integrates HPO *phenolyzer*[73] and EHR text mining to provide prioritization of causal genes for a patient[87]. This framework extracts phenotypes from the patient's narrative reports, sends them automatically to *Phenolyzer* and displays a list of potential causal genes.
Creating machine learning methods on EHRs data to perform general predictive models is difficult because of the quantity, high dimensionality, sparsity and quality of data [88]. Most of the time, machine-learning algorithms are centered on a specific pathological condition. They can be trained on literature and knowledge extracted from unstructured EHRs (text mining and PheWAS) then applied to EHRs' data to help clinicians diagnose a specific disease. Examples include Hunter syndrome[89], cardiac amyloidosis[90], and Allergic Bronchopulmonary Aspergillosis[91].
Liang et al. developed a machine-learning model to predict the primary diagnosis of a patient's encounter among a set of 50 diseases[88] by using the clinical data produced during the encounter. Their model was trained on a set of health reports manually annotated by pediatricians and obtained high diagnostic accuracy across multiple organ systems. However none of those diseases were rare.

Imaging and diagnosis

Regarding common diseases, several studies demonstrated strong performance in image-based diagnoses, notably in imaging[92], dermatology [93], and ophthalmology [94].
In the context of rare diseases, Gurovich et al. published deep learning methods on pictures of patients to detect facial dysmorphism associated with syndromic conditions[66]. Their system, called DeepGestalt, was trained on a dataset of over 17,000 images representing more than 200 syndromes. DeepGestalt is a form of next-generation phenotyping technology no longer based on the text but now on photographs.

Artificial Intelligence for care and management

The figure 3 detailes the different steps for developing and implementing a CDSS in CKD. Machine learning methods are also used to optimize care processes, for example anemia management in end-stage renal disease[95,96], prediction of response after neo-adjuvant chemoradiation[97], automated alerts on acute kidney injury[98] or outcome predictions of chronic kidney disease with random forest regression[99] or Support Vector Machine[100]. For this random forest regression, authors built the model using the data of 61,740 EHRs to predict kidney function and provide clinical decision support with high macro-averaged and micro-averaged metrics. Ennis et al showed that CDSS improves primary care physicians adherence to guidelines for laboratory monitoring of CKD[101].
Wang et al. decribed a protocol from the development of a regional CKD surveillance system to a CDSS that will be implemented into the hospital information system[102].

CDSS must be evaluated in a clinical context such as medical devices. A research in clinicaltrial.org with these keywords “kidney failure” and “ machine learning” finds 8 clinical trials (October 2019). Among these 8 trials, two are completed, two are active but not recruiting and four are recruiting patients. The query “kidney failure” and “clinical decision support” and “EHR” displays 9 studies which six were published after January 2018. The query (“rare diseases” or “rare disease”) and (“clinical decision support” or “machine learning”) and “EHR” does not find accurate study. These numbers illustrate how new is this technology in the clinician practice.

Clinical and omics data

Chen et al.[103] developed an algorithm to identify patient with ciliopathies. This group of rare and severe diseases are cause by ciliary dysfunction, associated to over 200 genes. Ciliopathies can affect all organs, and are divided into 30 rare syndromes whose clinical signs overlap each other. The authors addressed two tasks: (1) identifying patient with ciliopathies from patients with other nephropaties, and (2) phenotypic subtyping ciliopathies. For the first task, they compute similarity metrics on the phenotypes extracted from the EHR of Necker hospital. They evaluate several similarity metrics on 77 patients with ciliopathies amongst 10,462 patients with nephropathies. Unsupervised machine learning methods were applied to select the relevant phenotypes to accomplish this task: linear support vector selection with $l2$ penalty, tree-based selection and random forest selection. The linear support vector selection gave the best accuracy to find patients with ciliopathies with a precision of 76% in the 30 patients most similar to an index patient. For the second task they considered the clinical research data collected from 1,031 patients mutated on genes associated to ciliar dysfunction. The unsupervised clustering method showed strong concordance with expert knowledge. Similarity metrics applied to rare genetic disease in nephrology offer new perspective to recruit patients for research and reduce the diagnostic journey. The phenotypic similarity associated with omics data will address precision medicine, especially in the context of

complex rare diseases like ciliopathies. It is crucial to diagnose the patient before onset of renal failure, so that the patient can benefit from a drug as soon as possible.

EHRs and the hospital's collective memory

Several situations may require having access to the hospital memory to make an informed decision for a patient: (i) complex patients with rare conditions for whom the published treatment guidelines do not provide a clear recommendation, (ii) undiagnosed patients, (iii) finding undiagnosed patients eligible for diagnoses based on the presentation of index patients with new mutations.
At Stanford, a specialty consultation service composed by a team of medical and informatics experts mine EHRs data and other health databases to "learn from patients like mine". They provide a summary of what happens to patients similar to the patient proposed to the staff [104].
Even with a search engine that enables the mining of historical data, it can be a complex task to find patients similar to an index patient. It may be difficult to choose which phenotypes must be used in the query. At the Imagine institute and Necker Hospital, a "patient like mine" tool is integrated in the data warehouse user interface. It computes a similarity distance between an index patient and all the patients of the data warehouse based on medical concepts extracted from the clinical text reports in the EHR. This tool provides the top20 most similar patients. If the index patient is undiagnosed, the clinician may use the information from the similar patients to reorient their clinical investigation. If the index patient is diagnosed, the clinician may find similar undiagnosed patients eligible for the same mutation[105]. The absence of a learning step is an advantage of this unsupervised method. In the particular context of rare diseases the low number of cases limits the usage of methods like deep learning approaches which require a large number of examples during the training phase.

Figure 4 summarizes how decision support systems for diagnosis are developed.

Toward a Learning Health System

The Institute of Medicine defined the learning health system in 2007[106]. The aim is to conduct effective translational research by accelerating the repurposing of clinical data for research and integrating the new knowledge into patient care. This convergence of the deployment of EHRs and the desire to accelerate translational research and reduce costs has allowed for the development of new areas of expertise in medical informatics[107].
The core objective is to learn from the collective experience of care delivery as recorded in the observational data. As seen previously in this article, much recent research has explored machine learning, data mining and data visualization methodologies to enhance personalized medicine with collective experience[108]. Such insights are referred to as Practice Based Evidence, or Real World Evidence[109].

Several initiatives are trying to move out of the local hospital setting to provide inter-hospital interoperability for learning health systems. The TRANSFoRm project promoted an architecture for the learning health system in Europe by using ontologies and archetypes[110]. A Learning Health System is particularly important in the context of rare diseases where each patient is a source of new knowledge and can help to reduce the diagnosis odyssey of the other patients. PEDSnet is a national US network for children with inflammatory bowel disease and embeds several tools to accelerate translational research and precision medicine. In addition to clinicians and researchers, they extend the learning health system to communities of patients for the "common purpose of improving the health and lives of children"[111]. The Undiagnosed Disease Network[112] is also a national US network for undiagnosed patients. It provides a platform to share data, tools, and common protocols to accelerate the diagnosis of rare or previously unrecognized diseases.
Furthermore Blizinsky et al. developed a framework to take into account the variability of population, in particular the ancestral, social and environmental factors that must be included in the genomic-enabled learning health system. They highlight the potentiality of learning health system to improve equity of care in the age of precision medicine[113].

Issues and potential solutions

Several obstacles must be overcome to make routine data suitable for research, including data fragmentation[114], data quality, and data privacy[115]. As seen before, a solution commonly adopted to address data fragmentation is the development of data warehouses. Other obstacles are related to the quality of routine care data. While some clinicians complain about spending too much time entering data in legacy systems[116], they believe that EHR is essential for the quality of care,[117] and for personalized medicine [118]. Data quality can be assessed against a number of dimensions: completeness, validity, coherence and comparability, accessibility, usefulness, timeliness, prevention of duplicate records[119]. Research data are expected to be high quality data with respect to coherence and validity but with narrow scope and limited follow up. Conversely, EHR data exhibit strengths like extremely broad coverage and long follow-up; in addition, redundancy among data types can be exploited to check the consistency of phenotypic data.[120,121]. However, data processing steps are required when relevant information is present only in unstructured format, e.g., images or text reports, [122] to transform the entire EHR into a form suitable as input for a learning health system[123].
As for accessibility, another dimension of data quality, data sharing for research purposes has raised many concerns about privacy. Regarding data governance, some authors have claimed that some ambitious overly theoretical frameworks may fit neither practical needs[124] nor patients' vision since patients who have rare diseases, and life-threatening conditions may be more open to the concept of data reuse for research [125]. Part of the solution for trust-fabric is technical: all institutions have been leveraging their IT infrastructure to implement state-of-the-art functionalities regarding data access and security. The other part of the solution relies on ethics.

Ethics

After explaining the benefits of repurposing the EHR, and discussing building solidarity in the community ("the good of all of us is good for each of us"), Lee listed three recommendations to minimize risk and maximize benefits: (i) ensuring that data and results are valid, (ii) implementing data protection to reduce the risk of unauthorized disclosure of personal data, (iii) applying severe repercussions to bad actors responsible for data breaches[126].
In addition, she insists on the importance of training clinicians and researchers on proper usage of health data. Segal et al. adds the adoption of professional guidelines that will explain the decision support system in diagnosis[127]. Some physicians are also concerned about the use of their narratives, some of them use “shadow charts” to represent their semiological reasoning without it being used against them in a lawsuit.

Conclusion

Artificial intelligence based on EHRs to diagnose rare diseases is not yet well developed. Several obstacles must be overcome to accelerate the development of these very promising algorithms to reduce diagnostic wandering and improve patient care management. For example, many efforts must be made to efficiently analyze the free text in order to extract reliable and accurate information. To achieve this goal, we have to decompartmentalize care and research in order to accelerate the repurposing of clinical data for new knowledge and the repurposing of new knowledge for better care. We have described several algorithms developed during research projects but rarely integrated into the hospital information system for use by clinicians. Expectations of patients with rare disease and clinicians are high; learning health systems must strive to meet them.

DISCLOSURE:

The authors declare no conflict of interest.

Acknowledgments

We deeply thank Jan Niemira for his proofreading of the manuscript.
Nicolas Garcelon was partially financed by The French National Research Agency, under the C’IL-LICO project (17-RHUS-0002).

Figures:
Figure 1: Query designed for PubMed review
Figure 2: Challenges to developing a clinical decision support system, and knocking down the technical barriers
Figure 3: Steps to develop and implement CDSS in nephrology department.

Figure 4: Decision support system on EHR. From the patient EHR to the different results available

References

1. Toll E. The Cost of Technology. *JAMA* 2012; **307**: 2497–2498.
2. Sinsky C, Colligan L, Li L *et al.* Allocation of Physician Time in Ambulatory Practice: A Time and Motion Study in 4 Specialties. *Annals of Internal Medicine* 2016; **165**: 753.
3. Jamoom E, Beatty P, Bercovitz A *et al.* Physician adoption of electronic health record systems: United States, 2011. *NCHS Data Brief* 2012: 1–8.
4. Carter JH ed. *Electronic Health Records: A Guide for Clinicians and Administrators*. 2nd Revised edition edition. Philadelphia: American College of Physicians; 2008.
5. Frankovich J, Longhurst CA, Sutherland SM. Evidence-based medicine in the EMR era. *N. Engl. J. Med.* 2011; **365**: 1758–1759.
6. Wartman SA. Toward a virtuous cycle: the changing face of academic health centers. *Acad Med* 2008; **83**: 797–799.
7. The Shire Rare Disease Impact Report (2013 – US and UK population).
8. Shortliffe EH, Sepúlveda MJ. Clinical Decision Support in the Era of Artificial Intelligence. *JAMA* 2018; **320**: 2199–2200.
9. Jensen PB, Jensen LJ, Brunak S. Mining electronic health records: towards better research applications and clinical care. *Nat. Rev. Genet.* 2012; **13**: 395–405.
10. Kirby JC, Speltz P, Rasmussen LV *et al.* PheKB: a catalog and workflow for creating electronic phenotype algorithms for transportability. *J Am Med Inform Assoc* 2016; **23**: 1046–1052.
11. Fung KW, Richesson R, Bodenreider O. Coverage of Rare Disease Names in Standard Terminologies and Implications for Patients, Providers, and Research. *AMIA Annu Symp Proc* 2014; **2014**: 564–572.
12. Bearryman E. Does Your Rare Disease Have a Code? *EURORDIS* 2015.
13. Geva A, Gronsbell JL, Cai T *et al.* A Computable Phenotype Improves Cohort Ascertainment in a Pediatric Pulmonary Hypertension Registry. *J. Pediatr.* 2017; **188**: 224-231.e5.
14. Schreiweis B, Trinczek B, Köpcke F *et al.* Comparison of electronic health record system functionalities to support the patient recruitment process in clinical trials. *Int J Med Inform* 2014; **83**: 860–868.
15. Vawdrey DK, Weng C, Herion D *et al.* Enhancing electronic health records to support clinical research. *AMIA Jt Summits Transl Sci Proc* 2014; **2014**: 102–108.
16. Mascia C, Uva P, Leo S *et al.* OpenEHR modeling for genomics in clinical practice. *Int J Med Inform* 2018; **120**: 147–156.
17. Abrahão MTF, Nobre MRC, Gutierrez MA. A method for cohort selection of cardiovascular disease records from an electronic health record system. *Int J Med Inform* 2017; **102**: 138–149.
18. Bremond-Gignac D, Lewandowski E, Copin H. Contribution of Electronic Medical

Records to the Management of Rare Diseases. *Biomed Res Int* 2015; **2015**: 954283.
19. Santoro JD, Sandoval A, Ruzhnikov M *et al.* Use of electronic medical record templates improves quality of care for patients with infantile spasms. *Health Inf Manag* 2018: 1833358318794501.
20. Choquet R, Maaroufi M, de Carrara A *et al.* A methodology for a minimum data set for rare diseases to support national centers of excellence for healthcare and research. *J Am Med Inform Assoc* 2015; **22**: 76–85.
21. Köhler S, Vasilevsky NA, Engelstad M *et al.* The Human Phenotype Ontology in 2017. *Nucleic Acids Res* 2017; **45**: D865–D876.
22. Home - OMIM - NCBI.
23. INSERM. Orphadata: Free access data from Orphanet. © INSERM 1997. Available on http://www.orphadata.org. Data version (XML data version). 1997.
24. Bender D, Sartipi K. HL7 FHIR: An Agile and RESTful approach to healthcare information exchange. In: *Proceedings of the 26th IEEE International Symposium on Computer-Based Medical Systems*. 2013, pp.326–331.
25. Chelsom J, Dogar N. Linking Health Records with Knowledge Sources Using OWL and RDF. *Stud Health Technol Inform* 2019; **257**: 53–58.
26. Storf H, Schaaf J, Kadioglu D *et al.* [Registries for rare diseases : OSSE - An open-source framework for technical implementation]. *Bundesgesundheitsblatt Gesundheitsforschung Gesundheitsschutz* 2017; **60**: 523–531.
27. Faria-Campos AC, Hanke LA, Batista PHS *et al.* An innovative electronic health records system for rare and complex diseases. *BMC Bioinformatics* 2015; **16 Suppl 19**: S4.
28. Wang AY, Sable JH, Spackman KA. The SNOMED clinical terms development process: refinement and analysis of content. *Proc AMIA Symp* 2002: 845–849.
29. Weng C, Bigger JT, Busacca L *et al.* Comparing the effectiveness of a clinical registry and a clinical data warehouse for supporting clinical trial recruitment: a case study. *AMIA Annu Symp Proc* 2010; **2010**: 867–871.
30. Lowe HJ, Ferris TA, Hernandez PM *et al.* STRIDE--An integrated standards-based translational research informatics platform. *AMIA Annu Symp Proc* 2009; **2009**: 391–395.
31. Zhou X, Chen S, Liu B *et al.* Development of traditional Chinese medicine clinical data warehouse for medical knowledge discovery and decision support. *Artificial Intelligence in Medicine* 2010; **48**: 139–152.
32. Krasowski MD, Schriever A, Mathur G *et al.* Use of a data warehouse at an academic medical center for clinical pathology quality improvement, education, and research. *J Pathol Inform* 2015; **6**: 45.
33. Kortüm KU, Müller M, Kern C *et al.* Using electronic health records to build an ophthalmological data warehouse and visualize patients' data. *Am. J. Ophthalmol.* 2017.
34. Gipson DS, Selewski DT, Massengill SF *et al.* NephCure Accelerating Cures Institute: A Multidisciplinary Consortium to Improve Care for Nephrotic Syndrome. *Kidney Int Rep* 2018; **3**: 439–446.
35. Murphy SN, Mendis ME, Berkowitz DA *et al.* Integration of clinical and genetic data in the i2b2 architecture. *AMIA Annu Symp Proc* 2006; **2006**: 1040.

36. Murphy SN, Weber G, Mendis M *et al.* Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2). *J Am Med Inform Assoc* 2010; **17**: 124–130.
37. Garcelon N, Neuraz A, Salomon R *et al.* A clinician friendly data warehouse oriented toward narrative reports: Dr Warehouse. *J Biomed Inform* 2018.
38. Davis MF, Sriram S, Bush WS *et al.* Automated extraction of clinical traits of multiple sclerosis in electronic medical records. *J Am Med Inform Assoc* 2013; **20**: e334–e340.
39. Software – OHDSI.
40. Hripcsak G, Duke JD, Shah NH *et al.* Observational Health Data Sciences and Informatics (OHDSI): Opportunities for Observational Researchers. *Stud Health Technol Inform* 2015; **216**: 574–578.
41. Cuggia M, Bayat S, Garcelon N *et al.* A full-text information retrieval system for an epidemiological registry. *Stud Health Technol Inform* 2010; **160**: 491–495.
42. Raghavan P, Chen JL, Fosler-Lussier E *et al.* How essential are unstructured clinical narratives and information fusion to clinical trial recruitment? *AMIA Jt Summits Transl Sci Proc* 2014; **2014**: 218–223.
43. Escudié J-B, Rance B, Malamut G *et al.* A novel data-driven workflow combining literature and electronic health records to estimate comorbidities burden for a specific disease: a case study on autoimmune comorbidities in patients with celiac disease. *BMC Med Inform Decis Mak* 2017; **17**: 140.
44. Garcelon N, Neuraz A, Benoit V *et al.* Improving a full-text search engine: the importance of negation detection and family history context to identify cases in a biomedical data warehouse. *J Am Med Inform Assoc* 2016.
45. Friedlin J, McDonald CJ. Using A Natural Language Processing System to Extract and Code Family History Data from Admission Reports. *AMIA Annu Symp Proc* 2006; **2006**: 925.
46. Goryachev S, Sordo M, Zeng Q *et al.* Implementation and evaluation of four different methods of negation detection. *Boston, MA: DSG* 2006.
47. Chapman WW, Hillert D, Velupillai S *et al.* Extending the NegEx lexicon for multiple languages. *Stud Health Technol Inform* 2013; **192**: 677–681.
48. Bill R, Pakhomov S, Chen ES *et al.* Automated extraction of family history information from clinical notes. *AMIA Annu Symp Proc* 2014; **2014**: 1709–1717.
49. Mukherjee P, Leroy G, Kauchak D *et al.* NegAIT: A new parser for medical text simplification using morphological, sentential and double negation. *Journal of Biomedical Informatics* 2017; **69**: 55–62.
50. Ford E, Carroll JA, Smith HE *et al.* Extracting information from the text of electronic medical records to improve case detection: a systematic review. *J Am Med Inform Assoc* 2016; **23**: 1007–1015.
51. Schmidt D, Profitlich H-J, Sonntag D. Towards Integrated Information Extraction and Facetted Search Applications in Nephrology. In: *Joint Proceedings of the 2th Workshop on Emotions, Modality, Sentiment Analysis and the Semantic Web and the 1st International Workshop on Extraction and Processing of Rich Semantics from Medical Texts co-located with ESWC 2016,*.Vol 1613. Mauro Dragoni and Diego Reforgiato Recupero and Kerstin

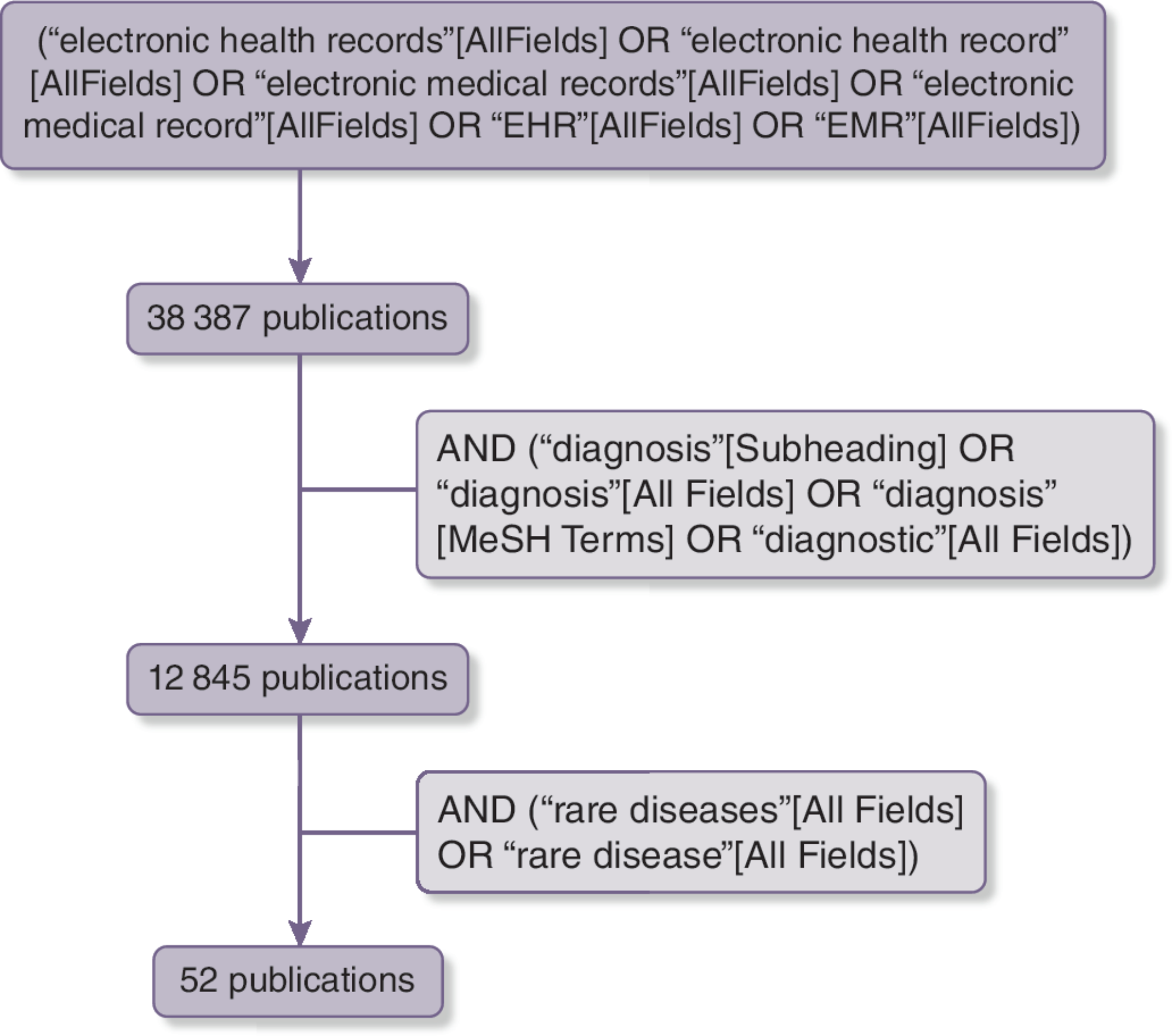
Denecke and Yihan Deng and Thierry Declerck. Heraklion, Greece: CEUR-WS.org, 2016.
52. Sonntag D, Profitlich H-J. An architecture of open-source tools to combine textual information extraction, faceted search and information visualisation. *Artificial Intelligence in Medicine* 2019; **93**: 13–28.
53. Hanauer DA, Mei Q, Law J *et al.* Supporting information retrieval from electronic health records: A report of University of Michigan's nine-year experience in developing and using the Electronic Medical Record Search Engine (EMERSE). *J Biomed Inform* 2015; **55**: 290–300.
54. Kohane IS. Using electronic health records to drive discovery in disease genomics. *Nat. Rev. Genet.* 2011; **12**: 417–428.
55. Savova GK, Masanz JJ, Ogren PV *et al.* Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. *J Am Med Inform Assoc* 2010; **17**: 507–513.
56. Métivier J-P, Serrano L, Charnois T *et al.* Automatic Symptom Extraction from Texts to Enhance Knowledge Discovery on Rare Diseases. In: Holmes JH, Bellazzi R, Sacchi L, et al., eds. *Artificial Intelligence in Medicine*. Springer International Publishing, 2015, pp.249–254.
57. Wei W-Q, Denny JC. Extracting research-quality phenotypes from electronic health records to support precision medicine. *Genome Medicine* 2015; **7**: 41.
58. Friedman C. Towards a comprehensive medical language processing system: methods and issues. *Proc AMIA Annu Fall Symp* 1997: 595–599.
59. Wu Y, Xu J, Jiang M *et al.* A Study of Neural Word Embeddings for Named Entity Recognition in Clinical Text. *AMIA Annu Symp Proc* 2015; **2015**: 1326–1333.
60. Adamusiak T, Shimoyama N, Shimoyama M. Next Generation Phenotyping Using the Unified Medical Language System. *JMIR Med Inform* 2014; **2**.
61. Holmes AB, Hawson A, Liu F *et al.* Discovering disease associations by integrating electronic clinical data and medical literature. *PLoS ONE* 2011; **6**: e21132.
62. Garcelon N, Neuraz A, Salomon R *et al.* Next generation phenotyping using narrative reports in a rare disease clinical data warehouse. *Orphanet J Rare Dis* 2018; **13**: 85.
63. Shen F, Zhao Y, Wang L *et al.* Rare disease knowledge enrichment through a data-driven approach. *BMC Med Inform Decis Mak* 2019; **19**: 32.
64. Namjou B, Marsolo K, Caroll RJ *et al.* Phenome-wide association study (PheWAS) in EMR-linked pediatric cohorts, genetically links PLCL1 to speech language development and IL5-IL13 to Eosinophilic Esophagitis. *Front Genet* 2014; **5**: 401.
65. Hebbring SJ, Rastegar-Mojarad M, Ye Z *et al.* Application of clinical text data for phenome-wide association studies (PheWASs). *Bioinformatics* 2015.
66. Gurovich Y, Hanani Y, Bar O *et al.* Identifying facial phenotypes of genetic disorders using deep learning. *Nature Medicine* 2019; **25**: 60.
67. Marbach F, Rustad CF, Riess A *et al.* The Discovery of a LEMD2-Associated Nuclear Envelopathy with Early Progeroid Appearance Suggests Advanced Applications for AI-Driven Facial Phenotyping. *Am. J. Hum. Genet.* 2019.
68. Perer A, Wang F, Hu J. Mining and exploring care pathways from electronic medical

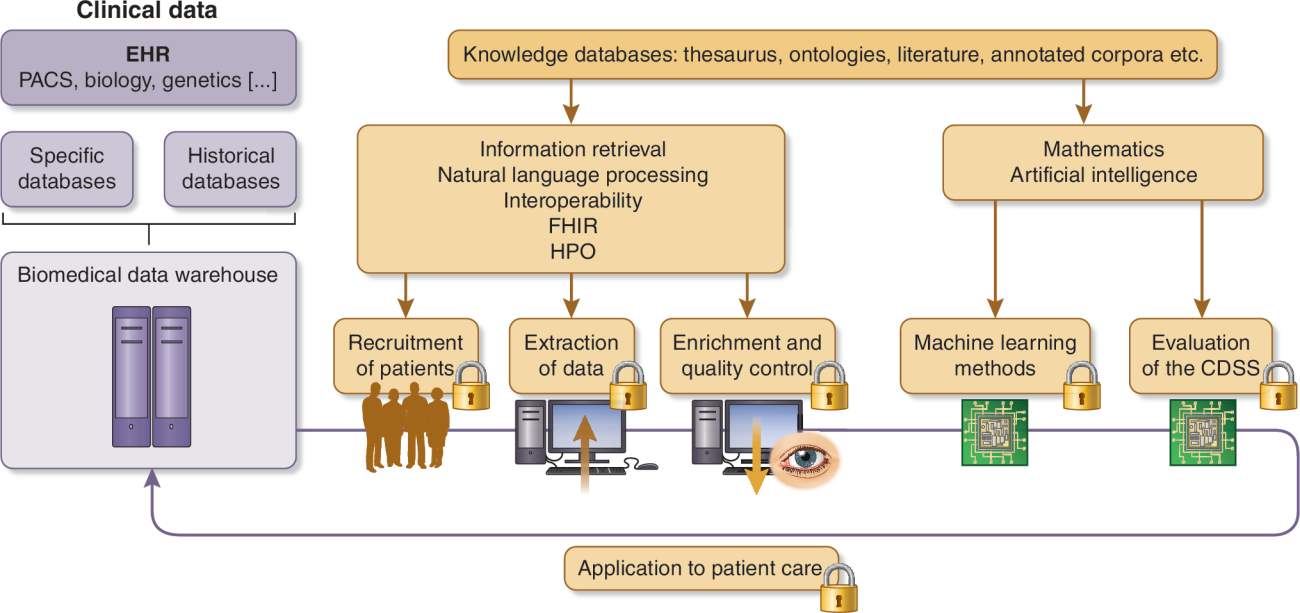
records with visual analytics. *Journal of Biomedical Informatics* 2015; **56**: 369–378.
69. Dagliati A, Sacchi L, Zambelli A *et al.* Temporal electronic phenotyping by mining careflows of breast cancer patients. *J Biomed Inform* 2017; **66**: 136–147.
70. Hurt L, Langley K, North K *et al.* Understanding and improving the care pathway for children with autism. *Int J Health Care Qual Assur* 2019; **32**: 208–223.
71. Zhang Y, Padman R. Innovations in chronic care delivery using data-driven clinical pathways. *Am J Manag Care* 2015; **21**: e661-668.
72. Dragusin R, Petcu P, Lioma C *et al.* FindZebra: a search engine for rare diseases. *Int J Med Inform* 2013; **82**: 528–538.
73. Yang H, Robinson PN, Wang K. Phenolyzer: phenotype-based prioritization of candidate genes for human diseases. *Nature Methods* 2015; **12**: 841–843.
74. Adler A, Kirchmeier P, Reinhard J *et al.* PhenoDis: a comprehensive database for phenotypic characterization of rare cardiac diseases. *Orphanet J Rare Dis* 2018; **13**: 22.
75. Müller T, Jerrentrup A, Schäfer JR. [Computer-assisted diagnosis of rare diseases]. *Internist (Berl)* 2018; **59**: 391–400.
76. Chen J, Xu H, Jegga A *et al.* Novel phenotype-disease matching tool for rare genetic diseases. *Genet. Med.* 2018.
77. Lindberg DA, Humphreys BL, McCray AT. The Unified Medical Language System. *Methods Inf Med* 1993; **32**: 281–291.
78. Jia J, An Z, Ming Y *et al.* eRAM: encyclopedia of rare disease annotations for precision medicine. *Nucleic Acids Res.* 2018; **46**: D937–D943.
79. Köhler S, Doelken SC, Ruef BJ *et al.* Construction and accessibility of a cross-species phenotype ontology along with gene annotations for biomedical research. *F1000Res* 2013; **2**: 30.
80. Ashburner M, Ball CA, Blake JA *et al.* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* 2000; **25**: 25–29.
81. Landrum MJ, Lee JM, Riley GR *et al.* ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res.* 2014; **42**: D980-985.
82. Li Q, Zhao K, Bustamante CD *et al.* Xrare: a machine learning method jointly modeling phenotypes and genetic evidence for rare disease diagnosis. *Genet. Med.* 2019.
83. Jia J, Wang R, An Z *et al.* RDAD: A Machine Learning System to Support Phenotype-Based Rare Disease Diagnosis. *Front Genet* 2018; **9**: 587.
84. Bastarache L, Hughey JJ, Hebbring S *et al.* Phenotype risk scores identify patients with unrecognized Mendelian disease patterns. *Science* 2018; **359**: 1233–1239.
85. Wulff A, Haarbrandt B, Tute E *et al.* An interoperable clinical decision-support system for early detection of SIRS in pediatric intensive care using openEHR. *Artificial Intelligence in Medicine* 2018; **89**: 10–23.
86. Simmons M, Singhal A, Lu Z. Text Mining for Precision Medicine: Bringing Structure to EHRs and Biomedical Literature to Understand Genes and Health. *Adv. Exp. Med. Biol.* 2016; **939**: 139–166.
87. Son JH, Xie G, Yuan C *et al.* Deep Phenotyping on Electronic Health Records Facilitates Genetic Diagnosis by Clinical Exomes. *The American Journal of Human Genetics* 2018; **103**:

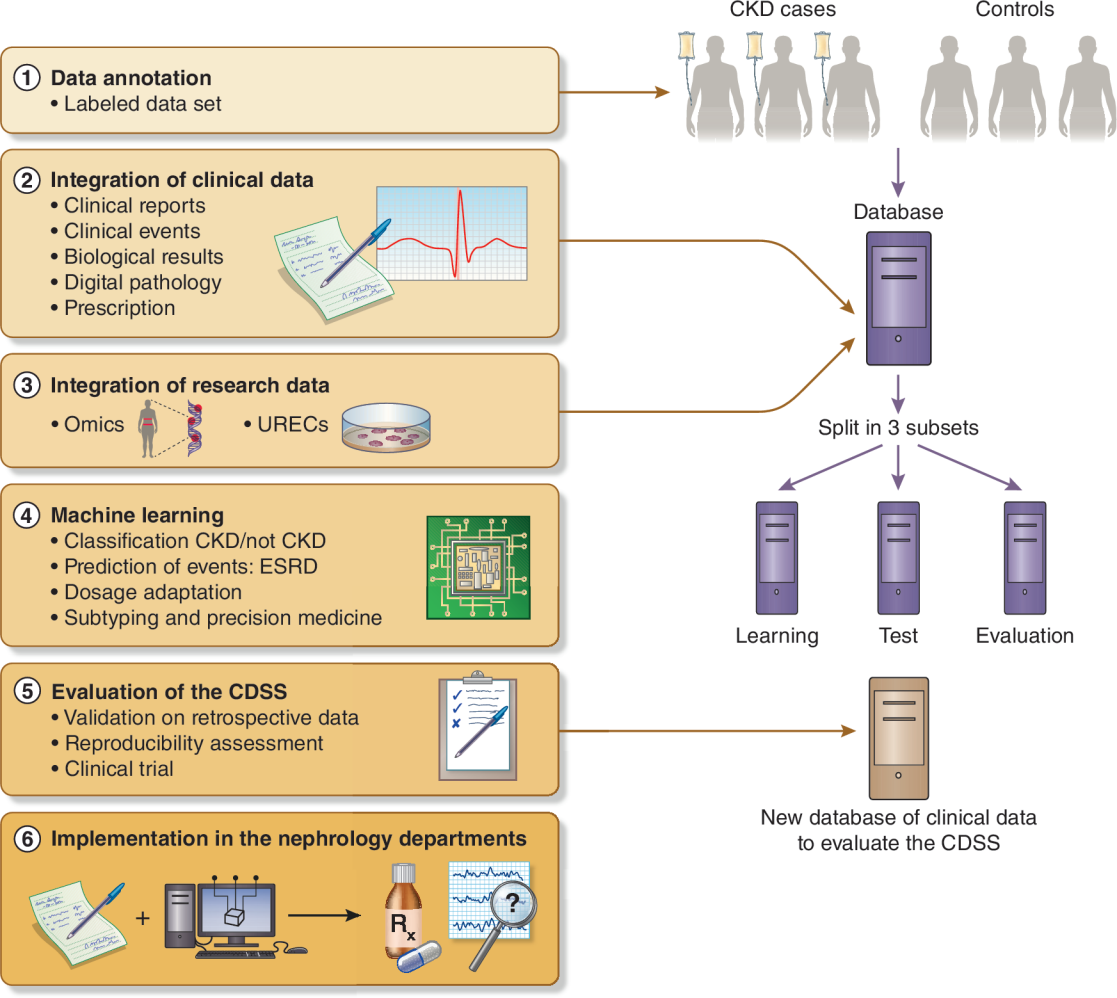
58–73.
88. Liang H, Tsui BY, Ni H *et al.* Evaluation and accurate diagnoses of pediatric diseases using artificial intelligence. *Nat. Med.* 2019.
89. Ehsani-Moghaddam B, Queenan JA, MacKenzie J *et al.* Mucopolysaccharidosis type II detection by Naïve Bayes Classifier: An example of patient classification for a rare disease using electronic medical records from the Canadian Primary Care Sentinel Surveillance Network. *PLoS ONE* 2018; **13**: e0209018.
90. Garg R, Dong S, Shah S *et al.* A Bootstrap Machine Learning Approach to Identify Rare Disease Patients from Electronic Health Records. *arXiv:1609.01586 [cs]* 2016.
91. Maguire A, Johnson ME, Denning DW *et al.* Identifying rare diseases using electronic medical records: the example of allergic bronchopulmonary aspergillosis. *Pharmacoepidemiol Drug Saf* 2017; **26**: 785–791.
92. Zimmer VA, Glocker B, Hahner N *et al.* Learning and combining image neighborhoods using random forests for neonatal brain disease classification. *Med Image Anal* 2017; **42**: 189–199.
93. Esteva A, Kuprel B, Novoa RA *et al.* Dermatologist-level classification of skin cancer with deep neural networks. *Nature* 2017; **542**: 115–118.
94. 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**: 2402–2410.
95. Brier ME, Gaweda AE. Artificial intelligence for optimal anemia management in end-stage renal disease. *Kidney Int.* 2016; **90**: 259–261.
96. Barbieri C, Molina M, Ponce P *et al.* An international observational study suggests that artificial intelligence for clinical decision support optimizes anemia management in hemodialysis patients. *Kidney Int.* 2016; **90**: 422–429.
97. Bibault J-E, Giraud P, Housset M *et al.* Deep Learning and Radiomics predict complete response after neo-adjuvant chemoradiation for locally advanced rectal cancer. *Sci Rep* 2018; **8**: 12611.
98. Kashani KB. Automated acute kidney injury alerts. *Kidney Int.* 2018; **94**: 484–490.
99. Zhao J, Gu S, McDermaid A. Predicting outcomes of chronic kidney disease from EMR data based on Random Forest Regression. *Math Biosci* 2019.
100. Polat H, Danaei Mehr H, Cetin A. Diagnosis of Chronic Kidney Disease Based on Support Vector Machine by Feature Selection Methods. *J Med Syst* 2017; **41**: 55.
101. Ennis J, Gillen D, Rubenstein A *et al.* Clinical decision support improves physician guideline adherence for laboratory monitoring of chronic kidney disease: a matched cohort study. *BMC Nephrol* 2015; **16**.
102. Wang J, Bao B, Shen P *et al.* Using electronic health record data to establish a chronic kidney disease surveillance system in China: protocol for the China Kidney Disease Network (CK-NET)-Yinzhou Study. *BMJ Open* 2019; **9**.
103. Chen X, Garcelon N, Neuraz A *et al.* Phenotypic similarity for rare disease: ciliopathy diagnoses and subtyping. *J Biomed Inform* 2019: 103308.
104. Gombar S, Callahan A, Califf R *et al.* It is time to learn from patients like mine. *npj*

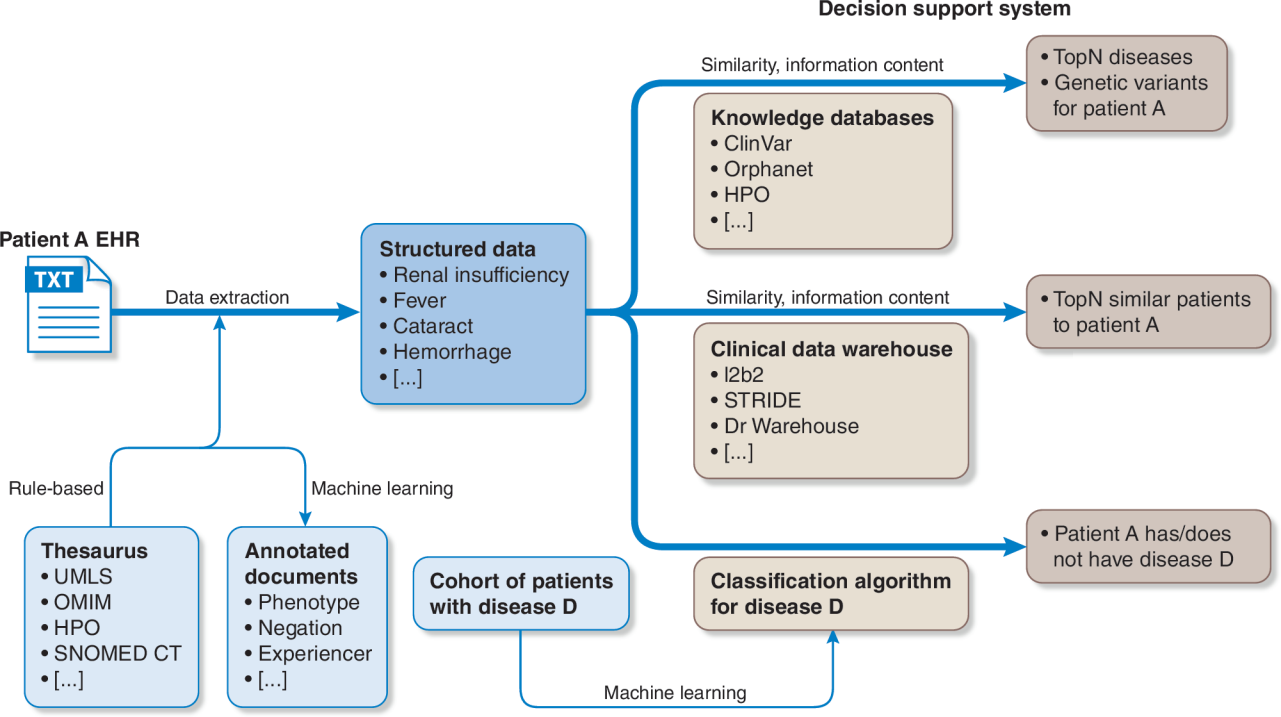
Digital Medicine 2019; **2**: 16.
105. Garcelon N, Neuraz A, Benoit V *et al.* Finding patients using similarity measures in a rare diseases-oriented clinical data warehouse: Dr. Warehouse and the needle in the needle stack. *Journal of Biomedical Informatics* 2017; **73**: 51–61.
106. Etheredge LM. A rapid-learning health system. *Health Aff (Millwood)* 2007; **26**: w107-118.
107. Lowes LP, Noritz GH, Newmeyer A *et al.* "Learn From Every Patient": implementation and early results of a learning health system. *Dev Med Child Neurol* 2017; **59**: 183–191.
108. Hu J, Perer A, Wang F. Data Driven Analytics for Personalized Healthcare. In: Weaver CA, Ball MJ, Kim GR, et al., eds. *Healthcare Information Management Systems*. Cham: Springer International Publishing, 2016, pp.529–554.
109. Smith M, Saunders R, Stuckhardt L *et al. A Continuously Learning Health Care System*. National Academies Press (US); 2013.
110. Delaney BC, Curcin V, Andreasson A *et al.* Translational Medicine and Patient Safety in Europe: TRANSFoRm--Architecture for the Learning Health System in Europe. *Biomed Res Int* 2015; **2015**: 961526.
111. Forrest CB, Margolis P, Seid M *et al.* PEDSnet: how a prototype pediatric learning health system is being expanded into a national network. *Health Aff (Millwood)* 2014; **33**: 1171–1177.
112. Ramoni RB, Mulvihill JJ, Adams DR *et al.* The Undiagnosed Diseases Network: Accelerating Discovery about Health and Disease. *Am J Hum Genet* 2017; **100**: 185–192.
113. Blizinsky KD, Bonham VL. Leveraging the Learning Health Care Model to Improve Equity in the Age of Genomic Medicine. *Learn Health Syst* 2018; **2**.
114. Ainsworth J, Buchan I. Combining Health Data Uses to Ignite Health System Learning. *Methods Inf Med* 2015; **54**: 479–487.
115. Holmes JH, Soualmia LF, Séroussi B. A 21st Century Embarrassment of Riches: The Balance Between Health Data Access, Usage, and Sharing. *Yearb Med Inform* 2018; **27**: 5–6.
116. March 2013 MF. Reducing the bureaucracy burden on the NHS. *Health Service Journal*.
117. Degoulet P. The Virtuous Circles of Clinical Information Systems: a Modern Utopia. *Yearb Med Inform* 2016: 256–263.
118. Armstrong S. Data, data everywhere: the challenges of personalised medicine. *BMJ* 2017; **359**: j4546.
119. Kodra Y, Posada de la Paz M, Coi A *et al.* Data Quality in Rare Diseases Registries. *Adv. Exp. Med. Biol.* 2017; **1031**: 149–164.
120. Neuraz A, Chouchana L, Malamut G *et al.* Phenome-wide association studies on a quantitative trait: application to TPMT enzyme activity and thiopurine therapy in pharmacogenomics. *PLoS Comput. Biol.* 2013; **9**: e1003405.
121. Wei W-Q, Teixeira PL, Mo H *et al.* Combining billing codes, clinical notes, and medications from electronic health records provides superior phenotyping performance. *J Am Med Inform Assoc* 2016; **23**: e20-27.
122. Escudié J-B, Rance B, Malamut G *et al.* A novel data-driven workflow combining literature and electronic health records to estimate comorbidities burden for a specific disease:

a case study on autoimmune comorbidities in patients with celiac disease. *BMC Med Inform Decis Mak* 2017; **17**: 140.
123. Garcelon N, Neuraz A, Benoit V *et al.* Improving a full-text search engine: the importance of negation detection and family history context to identify cases in a biomedical data warehouse. *J Am Med Inform Assoc* 2017; **24**: 607–613.
124. Lea NC, Nicholls J, Fitzpatrick NK. Between Scylla and Charybdis: Charting the Wicked Problem of Reusing Health Data for Clinical Research Informatics. *Yearb Med Inform* 2018; **27**: 170–176.
125. Rare Diseases: Why Diagnosis Can Be So Difficult - Canoe.com.
126. Lee LM. Ethics and subsequent use of electronic health record data. *J Biomed Inform* 2017; **71**: 143–146.
127. Segal MM, Rahm AK, Hulse NC *et al.* Experience with Integrating Diagnostic Decision Support Software with Electronic Health Records: Benefits versus Risks of Information Sharing. *EGEMS (Wash DC)* 2017; **5**.


("electronic health records"[AllFields] OR "electronic health record"[AllFields] OR "electronic medical records"[AllFields] OR "electronic medical record"[AllFields] OR "EHR"[AllFields] OR "EMR"[AllFields])
38 387 publications
AND ("diagnosis"[Subheading] OR "diagnosis"[All Fields] OR "diagnosis"[MeSH Terms] OR "diagnostic"[All Fields])
12 845 publications
AND ("rare diseases"[All Fields] OR "rare disease"[All Fields])
52 publications


Clinical data
EHR
PACS, biology, genetics [...]
Specific databases
Historical databases
Biomedical data warehouse
Knowledge databases: thesaurus, ontologies, literature, annotated corpora etc.
Information retrieval
Natural language processing
Interoperability
FHIR
HPO
Mathematics
Artificial intelligence
Recruitment of patients
Extraction of data
Enrichment and quality control
Machine learning methods
Evaluation of the CDSS
Application to patient care


CKD cases
Controls
1 Data annotation
• Labeled data set
2 Integration of clinical data
• Clinical reports
• Clinical events
• Biological results
• Digital pathology
• Prescription
Database
3 Integration of research data
• Omics
• URECs
Split in 3 subsets
4 Machine learning
• Classification CKD/not CKD
• Prediction of events: ESRD
• Dosage adaptation
• Subtyping and precision medicine
Learning
Test
Evaluation
5 Evaluation of the CDSS
• Validation on retrospective data
• Reproducibility assessment
• Clinical trial
New database of clinical data
to evaluate the CDSS
6 Implementation in the nephrology departments
Rx
?


Decision support system
Patient A EHR
TXT
Data extraction
Structured data
• Renal insufficiency
• Fever
• Cataract
• Hemorrhage
• [...]
Similarity, information content
Knowledge databases
• ClinVar
• Orphanet
• HPO
• [...]
• TopN diseases
• Genetic variants for patient A
Similarity, information content
Clinical data warehouse
• I2b2
• STRIDE
• Dr Warehouse
• [...]
• TopN similar patients to patient A
Classification algorithm for disease D
• Patient A has/does not have disease D
Rule-based
Machine learning
Thesaurus
• UMLS
• OMIM
• HPO
• SNOMED CT
• [...]
Annotated documents
• Phenotype
• Negation
• Experiencer
• [...]
Cohort of patients with disease D
Machine learning