Chapter

Electronic Medical Records and Machine Learning in Approaches to Drug Development

Ayaka Shinozaki

Abstract

Electronic medical records (EMRs) were primarily introduced as a digital health tool in hospitals to improve patient care, but over the past decade, research works have implemented EMR data in clinical trials and omics studies to increase translational potential in drug development. EMRs could help discover phenotype-genotype associations, enhance clinical trial protocols, automate adverse drug event detection and prevention, and accelerate precision medicine research. Although feasible, data mining in EMRs still faces challenges. Existing machine learning tools may help overcome these bottlenecks in EMR mining to unlock new approaches in drug development. This chapter will explore the role of EMRs in drug development while evaluating the viability and bottlenecks of their uses in data mining. This will include discussions on EMR usage in drug development while highlighting successful outcomes in oncology and exploring ML tools to complement and enhance EMR as a widely accepted drug-research source, a section on current clinical applications of EMRs, and a conclusion to summarize and imagine what a future drug research pipeline from EMR to patient treatment may look like.

Keywords: drug research and development, machine learning, AI, electronic medical records, EMR, EHR, NLP, deep learning, big data, data analysis, data-mining

1. Introduction

Advances in Artificial Intelligence methods have skyrocketed in the past decade, especially in the medical space where the impact of healthcare reaches individuals across a broad spectrum of communities. In particular, machine learning (ML) researchers have gained access to a large quantity of high quality medical data, aggregated by health providers as a result of implementing hospital management systems. A crucial element of these management systems is electronic medical records (EMRs), which are rich in valuable real world data on patient, clinical and genomic data. An EMR is a digitized record of a medical occurrence documented either during or after an encounter by a medical professional in a medical environment. For example, the results of a blood test administered at a hospital may be part of an EMR. Clinical notes taken by the doctor in a routine check-up at a local clinic are also included in the EMR. EMRs can come in the form of structured data such as drug

orders, medications, laboratory tests and diagnosis codes or unstructured data such as text-based clinical progress notes, radiology reports and pathology findings [1].

When EMRs are amalgamated to create a longitudinal overview of a specific patient, this larger unit of digitized records is called an electronic health record (EHR). Since EHRs contain historical data, they are used to track the health progression of patients over time. Although in some sources, the terms EMR and EHR are used interchangeably, or are sometimes referred to as the electronic patient record, for simplicity the above definitions are used here. Another digital record is the personal health record, which is the electronic medical data that the individual may choose to provide to the medical institutions or health providers, however issues of personal choice in volunteering data are beyond the scope of this chapter, so we do not consider the personal health record here.

Today, providers produce EMRs with the hope to provide a centralized source of medical data, which helps increase care coordination. With a standardized EMR system, if an individual decides to switch health providers, the medical data can seamlessly transfer to the new institutions. Furthermore, centralized medical data reduces duplication of records and identifies missing patient data, which reduces valuable time spent in clinical care. Compared to the traditional paperwork, EMRs significantly decreases disease identification time, making healthcare more time efficient and cost effective [2, 3]. In this sense, the EMRs improve quality of care.

In reality, there are issues in introducing EMRs into healthcare provider systems such as implementation and workflow disruptions. Implementation requires funding, necessary staff, and up to date digital technology. Institutions and geographic regions with ample resources will benefit from this implementation. However, for many smaller scale practices, implementation is not financially viable. For regions where institutions do not have access to technology that enables the production, storage and sharing of EMRs, this concept does not make sense. Furthermore, workflow is disrupted when clinicians and other medical professionals must alter their workflow in order to complete these documents. EMRs are notoriously unpopular in the medical community as it burdens professionals to constantly type on their computer instead of caring for their patients. Burdened professionals do not see the long term benefits and the reality in medical environments is that EMRs are primarily used for financial and administrative purposes. For example, although there are no global standards to what may be included in an EHR, it must always have billing codes, which are used for administrative purposes such as reimbursement or auditing reasons.

Despite these institutional challenges, EMRs are gaining traction in the biomedical space because there is potential to extract important biomedical conclusions from EMRs. As of December 2019, there are just under 2.1 million papers published on electronic medical records in drug development and research within google scholar [4]. Because EMRs are untapped and vast in quantity, researchers are particularly focused on testing ML methods on EMRs. EMRs also provide resources to carry out clinical trials at a lower cost and with reduced duration in terms of efficiency gained from automation and having better data sources. With a manual approach to identify and extract high value data, drug research on EMRs are not scalable and are extremely costly to employ domain experts for data extraction. The push for medical document digitization in conjunction with recent development in ML methods, such as natural language processing (NLP) that allows for machines to mimic human comprehension of written text, has allowed the outsourcing of these research tasks to machines and further facilitate drug research.

In the context of ML methods, EMRs pose problems such as how EMRs do not have a standardized formatting, how minorities could be underrepresented, and how EMRs contain human errors. Today in the healthcare space, EMRs exist in

abundance but were not originally created with a large scale data-mining vision [5]. Rather, providers replaced paper-work with electronic records to keep up with the technological pace of the 21st century. Such digitization of the traditional paper-work was done on an ad-hoc basis and many healthcare institutions independently regulate EMRs to create a highly heterogeneous data set [6, 7]. This heterogeneity makes data pre-processing for ML methods time consuming and financially costly if domain experts are required for this task. Another difficulty stems from the issue of institutions and geographic regions not having access to technology or financial resources to implement EMRs. The lack of EMRs in particular communities means those individuals are not electronically visible. In this sense, EMRs will not be able to sample certain populations in the world. These underrepresented populations will not have as much benefit from the biomedical success of EMRs as those represented in the sample populations, increasing the inequality of medical care. Lastly, basic human error in the EMRs will affect analysis performed on these data sets, if they are not corrected. In addition, the EMRs come from different institutions, which may enter their data differently. Without a standardized requirement for EMRs, some parts will be missing core information and the operation is not scalable.

1.1 Chapter overview

This introduction started with a brief discussion of what an EMR is and how we define it in the absence of international unifying standards. This chapter will now move on to an overview of how machine learning techniques, applied to EMRs, are influencing three key areas of biomedical research and drug discovery: (1) phenotype-genotype associations, (2) clinical trials, and (3) pharmacovigilance.

Firstly, we assess the impact of EMRs on making accurate phenotype-genotype associations, where physical traits are linked to specific locus in the genome. We then look at EMRs in the context of clinical drug trials and pharmacovigilance, which together amount to the tracking of a drug's efficacy and adverse side-effects both before and after it is licensed and used. Finally, a number of different case studies are looked at in detail, and we present a vision of how integrated EMRs and ML-driven EMR drug research could be implemented in the future.

2. EMRs and phenotype-genotype association research

Phenotype-genotype association is the correspondence between a person's genetic makeup—their genotype, and the observable characteristics or pathologies that are a product of their genetics interacting with the environment—their phenotype. In the medical space, researchers study phenotype-genotype associations because variations in the human genome affect how a person exhibits phenotypic traits, so to understand phenotype-genotype relations is to have biological insight into disease mechanisms. Furthermore, phenotype-genotype associations are important in drug discovery because phenotype targets are used to identify viable drug targets within the human genome and are needed to understand the chemistry of a potential drug within the human biology. Understanding phenotype-genotype associations has useful downstream applications in many fields including disease categorization, phenotype discovery, pharmacogenomics, drug–drug interaction (DDI), and adverse drug event (ADE) detection, and genome-wide and phenome-wide association studies [8].

Phenotype-genotype association research owes its foundation to the genome-wide association studies (GWAS) studies that were driven by the potential of

genetic variations modulating disease risks, expression and progression. Although the GWAS studies accumulated vast amount of genetic data, a remaining challenge is translating genetic markers to its associated phenotype [9, 10]. A high-throughput solution to such challenge is to harness phenotype data embedded in EMRs.

In a medical provider setting, clinical professionals observe phenotypes on a daily basis to diagnose diseases because phenotypic traits are manifestations of an individual's genome interacting with the environment. Such diagnosis is recorded extensively in EMRs, making them rich in phenotype-related data. Following the human genome project and the following development in sequencing whole genomes, EMRs can now feasibly link an individual's genome as part of their medical data.

However, linking genomic data to EHRs is not common in clinical practice. This is due to the combination of clinics offloading new sequencing technology to bioinformatics laboratories and the lack of infrastructure for integrating the processed genomic data into EHRs [11]. Unlike most clinical laboratory tests, genomic testing requires data curation during the bioinformatics pipelines. Therefore, when laboratories send genomic tests back to the original provider, the format or structure of that data may not be directly compatible with the local EHR system [12]. In 2016, laboratories were still physically mailing or faxing genomic reports in PDFs, which is a format that is extremely difficult for machines to read and interpret [12]. This clinical hurdle aside, in biomedical research this genomic inclusion in EHRs shows potential in secondary use as raw data from which to draw medically meaningful results [2, 12, 13]. Assuming that the EMR has adequate phenomic and genomic data on an individual, algorithms can translate raw data in EMRs to phenotype data, which in turn can be associated with the genomic data.

This section will focus on studies that cover phenotype-genotype research using EMRs that aims to advance drug research, with particular attention to the machine learning methods used in these cases. In a broad sense, this phenotype-genotype application of EMRs to drug research has two major tasks. First is to identify phenotypes contained in EMRs and second is to extract the phenotype to genotype associations.

One of the validated processes to identify phenotypic traits from EMRs is the use of standardized codes. Standardized codes have been designed for specific medical needs and are heavily used in the structured documentation in EMRs. When composing EMR's, medical professionals use an internationally standardized set of codes for reporting disease and health conditions called the International Classification of Disease (ICD) developed by the World Health Organization (WHO). For example, the ICD code may be a procedure code that indicates what medical procedures a patient has received during hospitalization or a disease code that specifies a clinician's diagnosis. Although standardized, the recorded ICD relies on a consistent interpretation of the ICD criteria for accuracy and relevancy, which will inevitably vary between clinicians, departments and institutions. However, researchers circumvent the larger issue of heterogenous EMR data types, which might range from character strings in clinical notes to matrices of pixels in radiology images, by focusing on these codes that are a standardized part of EHRs.

In the context of AI, using standardized codes is advantageous because they vastly reduce the set of possible inputs to any given machine learning algorithm. In practical terms, the data requires little pre-processing, since the codes already contain accurate and rich medical information described by domain experts. Computation becomes scalable as less pre-processing means less manual work involved, which is a necessity when extracting phenotypic data. Inevitably, there are a multitude of competing standards. As mentioned earlier, the ICD is consistently updated

in order to internationally keep track of morbidity and mortality statistics with its eleventh version being adopted and replacing previous revisions starting 1 January 2022 [14]. In addition to the ICD, the US government has designed the ICD Clinical Modification (ICD-CM), which is based upon ICD but tailored to the US healthcare market. The Clinical Classification Software for ICD-CM, developed by the US Agency for Healthcare Research and Quality, is a further development to the ICD-CM that regroups codes into clinically relevant categories. New standards do not have to be based upon existing ones, however. Phecodes is a standard specifically designed for biomedical research and to facilitate phenome-wide association studies, first published in 2010 [15, 16]. In 2017, these different sets of standardized EMR codes (ICD, ICD-CM, phecodes) were compared based on their ability to create correctly pair single nucleotide polymorphism (SNP), which is a nucleotide level genetic variation, to the corresponding phenotype, and it was found that the phecodes performed markedly better than the ICD based standards [15, 17]. It is perhaps not surprising that the phecodes performed best. Phecodes were developed for research purposes, whereas ICD and related standards are more focused on record keeping and streamlining the financial aspect of healthcare. These results illustrate how common EMR codes used in hospitals are not well designed for ML purposes. Although these codes are a convenient aspect within the context of diverse data from EMRs, care must be taken when designing algorithms, which repurpose the codes for phenotype extraction.

EMRs often contain a mixture of standardized codes and free-text. To improve upon methods that only consider codes, machine learning tools, largely based upon NLPs, have been developed to collect more phenotypic data from data sources beyond standardized codes such as textual clinical notes, textual discharge summaries and radiology reports [1, 18–21]. Liao et al. developed a multimodal automated phenotyping (MAP) algorithm to leverage both ICD codes and EMR textual narratives based on the Unified Medical Language System [18]. MAP is multimodal because it can extract entities such as ICDs, medical NLP concepts and healthcare utilization information related to a certain phenotype from both codes and free text. Using MAP, Liao et al. analyzed those entities by different latent mixture models to predict whether a patient had a certain phenotypic feature. Liao et al. ran the algorithm through a validation dataset that contained labelled data with one of 16 unique phenotypes to show that MAP can extract relevant and phenotype-specific entities at comparable accuracy to those identified by a manual approach (AUC-MAP = 0.943, AUC-manual = 0.941). Another example of successful high throughput method to extract phenotypes from EMRs is PheNorm, which harnesses standardized codes as training labels and does not require domain experts to label the training set, making the model highly scalable and cost effective for phenotype research [19]. In the face of the ML hype, it is naive to say that ML methods are superior and domain experts will become superfluous in the future. For example, Coquet et al. demonstrated the use of NLP methods and a Convolutional Neural Network (CNN) method to create word embeddings in clinical notes to automate clinical phenotyping of prostate cancer patients [20]. In this particular case, the phenotyping accuracy of CNN model (F-measure = 0.918) surpassed that of the rule-based model (F-measure = 0.897) [20] and the authors concluded that the mixture of both models can lead to even better precision and accuracy. These statistics in which the CNN model, which is a class of deep neural networks, outperformed the rule-based model, an example of human driven modelling where domain knowledge is needed, is indicative of the potential in ML methods but human expertise is still needed to attain even higher accuracy and precision.

The next stage after phenotype extraction is to create phenotype-genotype associations. In addition to the development of higher quality and more available

electronic medical records, EHRs can now be matched with biopsies stored in biobanks through patient-specific identifiers making it possible to study genetic and phenotypic data alongside clinical findings. Earlier studies focused on using statistical methods, such as the proof of concept study done by Denny et al. to develop a method to scan phenomic data for genetic associations using ICD billing codes [16]. Subsequent studies have shown the viability of using ML algorithms to understand phenotype-genotype associations using EMR sources with most of the papers published in the past year [22, 23]. Recently, deep learning gained popularity as an accurate framework at identifying phenotype-genotype associations [24]. Boudellioua et al. takes a deep neural network and developed an OpenSource phenotype-based tool called DeepPVP, which prioritizes potential causative variants from whole genome sequence data [25]. As another example, Zeng et al. used Bayesian network learning to extract epistatic interactions, which are gene-to-gene interactions that change exhibited phenotypic traits, that effect breast cancer patient survival on 1981 EHRs taken from the METABRIC dataset [26]. Their model learned SNP associations that effect breast cancer patient survival that agreed with domain knowledge from breast cancer oncologists [26]. Furthermore, unsupervised learning has also been recognized as a great tool to discover new phenotypes [27]. Stark et al. studied the unsupervised extraction of phenotypes from cancer clinical notes to use in association studies and reported success in finding new phenotype-genotype association hypothesis that are not published but plausible from a biological perspective [27]. Positive results form many recent studies demonstrates how deep learning shows promise in phenotype-genotype association extraction.

Such high performing machine learning on big data to create phenotype-genotype associations give hope to the future of personalized medicine, which is healthcare tailored to different variations in a genotypes. More basic biomedical research on phenotype-genotype associations opens possibilities for selecting best treatments and for studying drugs that come back with negative or adverse results. However, getting to such advanced levels of drug research is still on the horizon as there are still more challenges in finding phenotype-genotype associations.

As mentioned before, one of the major problems is that EMRs generally suffers from the difficulty in identification and correction of missing or mistaken data. In many cases, ML methods require large datasets and when EHRs are amalgamated from multiple sources, a high number of varying kinds of errors are carried over to the data set and therefore propagate through to the algorithms. Due to the high throughput of data in ML methods, there is a need for an automatic correction filter, or a complete work around the missing data. One solution to missing EMR data is to identify the missing phenotype data and correct it using a combination of bioinformatics and genomic data [28, 29]. Even with sparse numbers of high quality phenotypic or genotypic data, there has been studies that have successfully extracted phenotype-genotype information from EMR using semi-supervised, bulk phenotyping framework, and NLP-based machine learning techniques [24, 30, 31]. Another method to tackle missing data is to use a machine learning model to completely encompass the missing data as part of the training set and therefore accept the sparsity as part of the valid data [32]. Another solution is to acknowledge the missing data as a variable in the modelling of the algorithm and quantify its predicted effects on the final results and conclusion [33].

In summary, EMRs are a vital source of information in basic biomedical science, specifically for phenotype-genotype associations, and there is a trend to test ML methods on this untapped and vast data set to overcome the challenges EMRs face during data mining. The advantage of EMRs is that it can be mined for phenotypes and linked to genomic data. The section discussed different types of standardized codes used in EMRs, which are easy to pre-process for ML frameworks. Codes such

as ICDs, ICD-CM, and phecodes showed that they can successfully and conveniently identify phenotypes. However, standard codes used by providers were not intended for data-mining purposes and therefore see performance issues when they are used outside their primary objective, to identify phenotypes. To harness EMR data beyond codes, studies look at a mixture of ICDs and free text. In the context of phenotype identification, this blend of data sources showed high performance especially when using ML methods in conjunction with more rule-based methods that require domain expertise. Furthermore, this section discussed the strong viability of ML methods for phenotype-genotype association identification, with a trend toward using deep learning frameworks. EMR applications through ML methods still face the problem of missing or erroneous data, which may affect the subsequent biomedical conclusions. Further work is being done to combat the shortcomings discussed and overall, EMRs have proven to be a promising data source for phenotype-genotype related research.

3. EMR use in clinical trials

Clinical research informatics has emerged in the last 5–6 years as a new field of biomedical translational research, which revolves around using informatics methods to collect, store, process and analyze real-world clinical data to further biomedical research purposes. With the increasing availability of such electronic data and the development of analysis tools, EMRs can help decrease the cost and time of clinical trials by automating patient recruitment, extend randomized control trials and enhance retrospective cohort studies.

Clinical trials are a crucial stage in drug development to test for drug safety and efficacy. These trials are time consuming, labor intensive and costly to operate, and a significant bottleneck for many trials is insufficient patient enrollment [34]. However, by harnessing the data contained within EMRs, clinical trials can become more efficient by automating recruitment and having a more extensive view of medical data compared to the traditional manual search. Successful examples have shown that EMR mining for potential recruitment are more cost efficient and less time consuming than traditional methods [35, 36]. As a quantitative example, a study done in the US studied 31 EHR-driven analysis on drug-to-genome interactions and concluded that EHRs helped decrease the trial cost by 72% per subject and reduced the duration of the studies [13].

It is also possible to repurpose systems that already exist within a clinical setting to improve trial recruitment. A study conducted by Devoe et al. repurposed an already existing Best Practice Alert (BPA) system, which was originally intended to improve patient care by automating basic keyword searches on patient EHRs, to recruit potential trial participants for a COPD study [37, 38]. Devoe et al. directly compared the cost effectiveness of the BPA-driven screening to that of the traditionally manual method, namely the EMR Reporting Workbench method where clinicians customize a query through a platform in order to pull data from the EHR database, and concluded that BPA was four times faster at screening all patients and ultimately lead to a projected 442.5 h reduction over the course of the study.

A particularly interesting case of a commercial EMR product developed for research purposes used in a clinical setting is a platform called InSite. This Software as a Service platform was developed out of the Electronic Health Record for Clinical Research (EHR4CR) project (completed Spring 2016), which aimed to create a secure, robust and scalable platform used around Europe to create a network of safe and security-compliant real world data, which can be reuse to further clinical research [39]. International research groups and medical providers from multiple

countries developed this platform and intended for researchers to interact with hospital-based EHRs. A study by Claerhout et al. studied the feasibility of using InSite as a tool to estimate numbers of eligible participants for clinical trials at 24 European hospitals [40]. They studied the inclusion and exclusion (I/E) criteria of protocols from 23 trials across diverse therapeutic areas, including ABP 980 and trastuzumab for early breast cancer, a combination of cediranib and chemotherapy in relapsed ovarian, fallopian tube or epithelial cancer, and selumetinib in combination with docetaxel for metastatic lung cancer. These clinical trials were sponsored by various pharmaceutical companies [1] to represent key I/E criterion using terms included in the standard medical coding systems [2] [40]. It was found that a median of 55% of the I/E criteria can be translated to InSite queries using the standard medical coding systems to correctly identify potential trial patients. This result is promising as it shows the feasibility of translating the complex protocol criteria into machine-readable queries via an already existing platform.

This success of patient identification is attributed to how well defined the disease parameters are in the I/E criterion and whether its clinical concepts exactly match a query that the InSite platform can digest. Unfortunately, these queries do not contain easily accessible nor standardized temporal information on disease development such as the rapid progression of a tumor size or the timing at which an operation was carried out. This lack of temporal resolution led to the lowest formalization rate (38%) in patients with metastatic melanoma, revealing the difficulty of acquiring temporal information on tumor staging and genetic testing [40]. A possible next step to this study is to harness NLP to the unstructured EMR data and to resolve the temporal issue in order to increase performance in patient recruitment. Overall, this study showed the potential for this commercialized platform for optimizing recruitment by hospitals. Beyond the feasibility of estimating the number of potential trial patients, this platform is advantageous because InSite offers a convenient and efficient way for researchers can access real-time clinical data by extracting relevant EMRs without disrupting healthcare providers with new technological implementations.

It has been shown that NLP [34] is able to reduce the amount of manual-driven patient identification required. Once the number of patients eligible for a clinical trial is estimated, the next step is to carry out patient screening on each individual. There are three methods that can carry out these checks. Meystre et al. harnessed NLP to directly compare clinical trial screen accuracy between machine learning, rule-based and cosine-similarity based methods and reported the highest accuracy (micro-averaged recall 90.9%) and precision (89.7%) for the machine learning method [34]. In such automations, the usage of NLP and harnessing machine learning is key to fully automating cohort selections using EHRs, and there are research done to further those tools, which is illustrated with the emergence of CREATE [41] and SemEHR, which is an open source semantic search and analysis tool for EMRs [42]. Such automations revolutionize clinical trial processes by cutting down administrative work by an order of magnitude. To deal with the ever increasing amount of EMR data made available, case studies have also shown that unsupervised ML methods may be used to identify disease cohort selection with high accuracy compared to the traditional and manual methods [43].

In some cases, EMRs can allow for more diversity in clinical trials and provide data collection on individuals that are traditionally underrepresented, such as racial minorities, children, rural communities or pregnant women [35, 44, 45]. However,

---

[1] Amgen, AstraZeneca, Bayer, Boehringer-Ingelheim, F-Hoffman La Roche, Janssen, Sanofi.

[2] Diagnosis: ICD-10CM, procedures: ICD-PCS, medication: ATC, laboratory: LOINC, clinical findings: SNOMED and anatomic pathology/oncology ICD-O-3.

there are also studies that published poor performance of information retrieval through EMR and ML [46]. There are high expectations for a new wave of ML tools to revolutionize medicine but researchers must be vigilant for unexpected biases arising from ML models trained on skewed or bad data.

For an example of bias in EMR driven selection of patients for trial, we look at the work of Aroda et al. They compared EMR-driven recruitment for type 2 diabetes patient across multiple health centers in the US to that of the traditional manual method [47]. Although Aroda et al. reported that the EMR-based recruitment had higher numbers of patients screening, better performance and improved randomizations, they also noticed an association with fewer women and racial minorities recruited. EMR and electronic-driven recruitment may cause bias in the type of cohorts identified, as electronically visible individuals are more likely to be identified and then consent to trials. A skew in this electronic visibility allow only certain cohort groups to be identified and studied in a clinical trial [48].

These biases arising from ML models are a significant aspect of drug research as they may cause inadvertent negative effects when these technologies are brought to market and into the medical centers. This may be the case of poor data sets or a poor selection of algorithms. In the real world, catch-all algorithms that work in academia sometimes fail and sometimes there is just not enough data for the data-hungry machine learning methods. Since manual methods do not suffer due to lack of scale when ML-based and data-driven research fail when they cannot access big data, the rise of ML driven processes will not make manual ones totally obsolete.

Another potential for EMR is to extend short, cost-limited trials by electronically monitoring the cohort after the trial is over. This creates a long term follow up without the cost associated with a traditional, extended clinical trial. There has been a successful case in testing novel probiotics to carry out a 5 year follow up, which would have been too expensive in traditional methods and retention rate increased due to this electronic method [49]. Furthermore, EMR data may be used in clinical trials beyond just a follow-up. There is interest in using EMRs as a primary data source or as a feasibility assessment tool in observational clinical trials, comparative effectiveness studies and randomized clinical trials [50]. In addition, data can be used to carry out retrospective cohort studies or population based cohort studies. Kibbelaar et al. proposed a method to combine data from population-based registries with detailed EHR to conduct an observational study and reported on a case study in an hemato-oncology randomized registry trial [51].

These implementations are dependent on the patient's consent to partake in the trials and there are studies that investigate the process and ethics of such consent [52]. Beskow et al. identified patient informed consent as a bottleneck in using EHR for randomized clinical trials. A study has also identified gaps in ethical responsibility in clinical studies carried out [53]. Furthermore, compliance to security and privacy regulations is a critical challenge as clinically produced EMRs proliferate through cloud platforms, mobile devices and commercialized technology. Whilst security and data protection are of paramount importance when dealing with EMRs, a discussion of the methods currently in use is beyond the scope of this chapter. The reader is directed to Refs. [54–56], in which the current technologies and methods used for security measures on EMRs are reviewed.

To conclude, using data within EMRs can help decrease the cost and time of clinical trials. First, the section discussed successful examples of EMR mining for potential recruitment in clinical trials, which included using systems that already exist in clinical settings, such as BPA and InSite, and tools that employ ML methods. An advantage with the use of ML methods in clinical trials is the increase in diversity in trial patients but there is still an issue with the bias that cause inequality in patient selection. Ultimately, the quality of the ML approach depends on the

quality of the training data. Therefore, with access to excellent data, EMRs can be used to extend short, financially limited trials or used as a primary data source to carry out aspects of data-driven clinical trials. Whilst ML methods are showing strong performance in enhancing clinical trials, big challenges remain before the data-driven method replaces the current clinical methodology.

4. EMR use in pharmacovalidation and data mining

However thoroughly a new drug is trialed and tested before it enters the market, it is possible that there are unknown adverse drug events (ADEs, colloquially known as side-effects) that manifest on time scales or in ways that cannot be seen in a clinical trial. Currently, adverse side effects of pharmaceutical products are a significant source for morbidity and are a significant healthcare cost in many countries [57, 58]. Therefore, it is vital that pharmaceutical companies undertake pharmacovigilance, in which they continually track the effects of their drugs after the drugs deployment. This means that clinical data on post-market drug effects has a high value to pharmaceutical companies [59]. Post-market surveillance of drugs to detect, evaluate and prevent ADEs with licensed drugs released in the market is called pharmacovigilance and is imperative for decreasing negative drug incidents.

Traditionally, medical professionals with domain knowledge would manually identify ADEs through sources such as clinical trials, health reports, published medical literature, observational literature and social media [60], which is time consuming and costly. Therefore, automatically mining these electronic narratives are an efficient way to identify negative events in the real world setting. Luckily, real world data on pharmaceutical products and their effects are richly logged in patient EHRs. To successfully mine the vast quantity of dense data in the EHRs for drug events, specifically ADEs, studies have focused on the narrative aspect of EMR and have successfully extracted ADE from both structured [61, 62] and unstructured [63–65] texts.

This focus on EHR narratives stems from studies that have shown that disease classification codes, such as ICD, used in EMRs do not encompass the symptoms, disease status and severity needed for ADE sensitivity and therefore are not appropriate in drug event mining [66–68]. Therefore it is necessary to extract more detailed information from the written text in EMRs, which is achieved using NLP algorithms. This is a two staged computational task. Firstly, the algorithm must perform accurate name entity recognition (NER) to identify diseases, drugs, and negative events in the text, and then it must quantify associations between those entities, to build a concept of what had occurred [69, 70].

Since 2012, significant developments in statistical analysis, machine-learning methods and heterogeneous data integration have allowed for automated ADE detection and offer tools for a novel, automated pharmacovigilance analytics [71]. Some statistical methods such as the odds ratio has been used by Leeper et al. and Banda et al. to create algorithms designed for extracting drug–ADE associations from EHRs [72, 73]. However, due to the need to define hypothesis using domain knowledge, experts in the field were necessary and this suggests a limitation that these statistical frameworks will not necessarily benefit from having more access to EHR resources because the core predictors depend on a priori knowledge, which is static within the algorithm. This means that there is currently still a manual element required in the process, which limits the scalability of this approach.

Some of the early EMR-narrative studies focused on keyword and phrase driven identification of general ADE. For example, there are semantic searches specializing in certain disease targets such as the work done by Ferrajolo et al. who looked at

drug related acute liver injury [74, 75] and Pathak et al. who mined for DDI between cardiovascular and gastroenterology pharmaceutical products [76, 77]. Although these disease specific searches may increase ADE detection in a certain medical domain, this tailored approach is not scalable or translatable to other diseases. In terms of identifying general ADEs without a target disease, Honigman et al. developed a search method using the Micromedex M2 D2 (Micromedex, Denver, Colorado) medical data dictionary to semantically associate drugs and drug classes to their negative effects and successfully showed the viability of keyword searches on EMRs [78, 79]. Chazard et al. went a step further to demonstrate searches on a variety of data structures such as drug administration records, laboratory results, and other clinical records to successfully detect general ADEs within free texts [80, 81]. These previous methods successfully identified general ADEs, but keyword driven searches are now considered simplistic and not scalable, but the success of even that method shows that there is great promise for modern techniques.

A further development to keyword-based semantics is a more symbolic rule-based search that looks for semantic patterns around drug and ADE entities. These symbolic rule-based searches allow for more information on dosage and non-standard terminologies to be identified during queries and are more capable of general ADE recognition [82–85]. With the rise of semantic research in the medical space, biomedical NER and NLP has been developed to aid clinical semantic searches and there are several open sources available, which have been adapted for ADE identification such as MedLEE [86], MetaMap [87], cTAKES [88, 89], MedEx [90], and GATE [91]. Of those, MedLEE and MetaMap are two of the most widely used, particularly in the pharmacovigilance space, where researchers extract Unified Medical Language System (UMLS) concepts from texts using NLP based approaches. Studies have shown the adaptability of these already available NLP systems. Banerjee et al. used grammar rules to extract all noun entities and then used MetaMap to semantically identify the type of entity found. This study found that medications are easily found as entities, but the model had difficulty in extracting symptoms from laboratory test results as they vary in length and word choices [92]. In adapting these NLP systems, each study hit limitations of each source and in particular these tools are not very capable in temporal resolution, which makes it difficult to distinguish drugs that cause ADEs from those products that indicate the presence of an ADE.

This shortcoming in temporal resolution has pushed for another wave of studies. In understanding the use of medication and mentions of diseases, the context surrounding these entities will determine whether the drug was or was not used at a time before or after an adverse incident. Some studies have created time stamps on event entities and medication administration in order to exclude situations where the adverse symptom was an already existing condition at drug administration, the ADE was due to another drug, the drug did not cause the ADE and is mentioned as a negative association, or the pharmaceutical product was given as treatment to the ADE [84, 93, 94]. Although time resolution on ADE events increase the accuracy of adverse incident detection, the vagueness and implicit tendency in the human language to describe temporal events remain as bottlenecks [95].

A great example to illustrate a collaborative ML research on clinical EMRs is the MADE1.0 challenge carried out in the US. This ML challenge illuminated the popularity and effectiveness of deep neural networking learning in identifying negative drug incidents, as these models counted for most submissions to the competition.

4.1 MADE1.0 challenge: pharmacovigilance on cancer patient EMRs

In the US, death due to a drug incidence is one of the top six causes of death with around 2–5% of hospitalized patients suffering from ADEs; in each case an adverse

event can increase healthcare cost by more than $3200 [96]. Traditionally, ADE-based pharmacovigilance is done by domain experts reading information on causality of drugs on incidents and temporal data on these events buried in the clinical narrative. However, this manual method is not scalable and very costly. To tackle the significant health and financial strain caused by ADEs, US research institutions participated in a machine learning challenge to develop methods automate real-time drug safety surveillance.

In 2018, University of Massachusetts (UMass) hosted a public NLP challenge to detect Medication and Adverse Drug Events from Electronic Health Records (MADE1.0). UMass provided 1092 longitudinal EHR notes, which were anonymized from 21 cancer patients from the University of Massachusetts Memorial Hospital. This EHR resource was rich with information on diseases, symptoms, indications, medications and relationships between these entities. Three main tasks were defined in this challenge: (1) named entity recognition (NER), which extracts drug medications, their attributes (dosage, drug administration, duration, etc.), disease indications, ADEs and severity, (2) relation identification (RI), which creates associations between entities, namely drug-indication, drug-ADE, and medication-attribute relations, and (3) the joint task that assess the NLP model's ability to perform both NER and RI. More detailed information on the challenge can be found at [96]. Jagannatha et al. reported that out of the 11 participating teams the highest F1 scores in each category was 0.8290 in NER, 0.8684 in RI, and 0.6170 in NER + RI, where the F1 score is the weighted mean of precision and recall with ranges from 0 (worst) up to 1 (best) [97].

Within NER task models, the main task can be distilled down to tokenizing sentences, so the tokens can then be labelled as specified entities. One common framework for NER is the hidden Markov model (HMM), in which the system is assumed to be the product of an unknown Markov process, which can then be statistically modelled. Conditional random fields (CRFs) are related to HMMs, however they differ in that, unlike HMMs, they are discriminative and classify labels by drawing decision boundaries. Unlike HMM, CRF does not have strict independence assumptions, which makes the model more flexible but highly complex at the training stage, meaning that retraining is more involved than that of the HMM [98]. The other main class of model is the neural network, including convolutional neural networks (CNN) and recurrent neural networks (RNN). Long short-term memory (LSTM) is an RNN architecture in common use for NER purposes. It is designed for classifications and predictions on time series data, in which events may occur with significant and unknown time lags in the sequence [99]. Teams involved in the MADE1.0 challenge used pre-trained embeddings to prepare the RNNs or as feature inputs into CRF training [97]. Within NER task models in this challenge, conditional random fields (CRF) and long short-term memory (LSTM) were among the most frequently used frameworks [97].

In the NER category, team WPI-Wunnava scored the highest scores with F1 = 0.8290 [97]. Wunnava et al. created a system called the Dual-Level Embeddings for Adverse Drug Event Detection (DLADE) to tailor to the NER task [100]. In the challenge, the NER task is limited to certain standard resources like NLTK, Stanford NLP, and cTakes for the text pre-processing for fairness of the participants with varying accessibility to resources. In particular, DLADE used training data and word embeddings provided by the challenge organizer as part of the publicly released resources. Wunnava et al. developed the system with a rule-based tokenizer, which first tokenized sentences, and then entities within sentences, where entities may be multiple words. The system then uses a combination of bi-LSTM, a model that examines the text sequence in the forward and reverse

direction to extract contextual representation, for the initial two layers responsible for the character embedding and the word embedding but employed a linear-chain CRF for the output layer [100]. Wunnava et al. concluded that their dual-level character and word embedding method was a better approach compared to the simple word-embeddings by showing a statistically significant ($p < 0.05$ and $p < 0.01$) improvement in F1-score over multiple entities (ADE, drug, dose, duration, etc.) [100]. However, many challenges remain when identifying multi-worded entities, unknown abbreviations, ambiguous differentiation between entities such as indication vs. ADE, and uses of colloquial or non-medical jargon.

In both the RI and NER-RI tasks, the process can be simplified to a classification problem, where entity pairs are in a certain class of relationships. Research teams used a variety of approaches to the RI tasks. As well as neural network methods, they also used random forest classifiers, in which an ensemble of decision trees is used and the aggregate score from the committee of decision trees decides the output class. Support vector machines (SVM) were another popular tool; they are optimizing algorithms that maximize the margin between the support vectors (input data) and the decision hyperplane [101].

In the RI category, team UofUtah-Patterson score the highest scores with F1 = 0.8684 [97]. Chapman et al. treated the RI task as a two-step supervised classification problem and employed random forest models implemented on scikit-learn to identify true relations between entities and to class the type of relation of the identified pair [102]. Their source code for their models submitted to the MADE1.0 challenge can be found on their github page [103] and details on the model architecture is authored at [104].

In the NER + RI category, team IBMResearch-dandala obtained the highest integrated task score (F1 = 0.6170) by harnessing bidirectional long short-term memory (BiLSTM) and CRF neural network for medical entity recognition, and a combined BiLSTM and attention network for relation extraction [97]. Dandala et al. reported that NER was achieved at high accuracy (F = 0.83) and RI measured an F score of 0.87 achieved by adding joint modelling techniques and using external resources as extra data inputs [105]. However high the individual F score, the overall integrated task only reached 0.6170, which suggests the need for domain knowledge to increase accuracy in ADE detection.

The MADE1.0 challenge highlights the potential for developing pharmacovalidation based on ML methods with very high performance in categories such as NER and RI, which are crucial in automated ADE extraction from EMRs. At the time of completion of the MADE1.0 challenge, Jagannatha et al. suggested two broad approaches to further improve the challenge's outcomes [97]. First, to work on designing methods that include external knowledge and unlabeled text, which suggests the potential for unsupervised learning. The second point was to increase efforts in higher volume, labelled corpus to train the models on, but this does not solve the issue of algorithms failing to adapt to the messy, real world EHRs, an inevitable encounter in commercial use. Not only did this challenge show success in developing ML-based pharmacovigilance but also demonstrated the power of collaboration and influenced other groups to further ADE research.

4.2 Further ML works and trends on pharmacovigilance

After the MADE1.0 challenge, an even further increase of available EHR resources has pushed researchers to develop robust ML methods, which are inherently data hungry and are predisposed to the vast amount of information provided by clinical texts. There is a study that builds on the MADE1.0 challenge and shows

the potential for deep learning models on EHR to extract ADE measures to help with pharmacovigilance. To try to solve the issue of under-reporting within the FDA Adverse Event Reporting System, Li et al. employed deep learning models and multi-task learning (MLT), in particular, hard parameter sharing, parameter regularization, and task relation learning, for ADE detection [106]. They used the MADE 1.0 challenge corpus, 1089 high-quality EMRs from oncology patients, for training and validation of their model. A BiLSTM conditional random field network was used for entity recognition and a BiLSTM-Attention network for entity relation extraction. Li et al. reported that the deep learning produced a F1 = 0.65 for the NER + RI task and this score was further improved through the hard parameter sharing MLT method to F1 = 0.67, whereas the other two MLTs did not improve performance. This study successfully built upon the findings from MADE1.0 and further improved the performance of the NER + RI task to show potential in this area.

Some ML trends that extract medically actionable results are the popularity of CRFs, SVMs, and random forest models. CRFs and SVMs may be used on languages beyond English. For example, Aramaki et al. studied Japanese clinical records and found that ADE were found in 7.7% of EHRs, out of which 59% can be automatically extracted [107]. They used CRFs and SVMs to determine whether a detected drug and adverse event pair was an ADE, which gave a 0.411 precision and 0.917 recall. In contrast, random forest models have been popular due to its reliable performance and explainability of the classifications when compared with other "black-box" models such as SVMs. Studies by Henriksson et al. and Wang et al. has used random forests for classification of entities and identify ADEs [108, 109]. Explainability of models is an often undervalued aspect of ML, but is valuable in the medical space. Overall, despite the many challenges, data-driven pharmacovigilance has advanced at an incredible pace owing to the mixture of funded challenges and developing ML methods and shows much promise to improve healthcare.

5. Drug repurposing

It is worth mentioning that EMR data can be mined for drug repurposing indications. The idea behind drug repurposing is to see whether existing, licensed drugs may have therapeutic benefits for conditions other than what they were designed for. Data-driven analysis is evidently key in this regard as it can detect drug response signals. Drug repurposing is different from the traditional drug discovery because data-driven analysis lacks a hypothesis for the indication intended to be treated or for the targeted biology. In other words, studies examine machine learning methods to see whether data-centric analysis can help create new hypothesis, which may either be a completely random and biologically impossible statement or a novel signal worthy of scientific investigation. Since drug repurposing only needs medical data and analytics, it is a cheap and quick alternative to the traditional drug discovery stages, which require basic research, pre-clinical research, clinical trials, and finally the review and approval of the pharmacogenomic product. The potential of drug repurposing is highly anticipated as this method requires big data and an increasing amount of digitized medical records such as EHRs are made available. It is a particularity popular topic in recent years as data-hungry machine learning tools develop and high-throughput server less machines are made cheaper and more accessible through cloud computing services such as AWS, Google Cloud Platform, and Microsoft Azure, to name a few. For a more in-depth discussion of oncology drug repurposing using data from EMRs, the reader is directed to Refs. [110–112].

6. Case studies in different countries

6.1 Oncology precision medicine in the US and Japan

Another anticipated but still young area is the possibility of precision medicine using individual genomic data. Cancer is an accumulation of genetic alternations within the cell and, oncogenetic or cancer-developing genes are called driver genes. Identifying driver genes within the genome and delivering the optimal treatment to such cancer-related targets is known as precision medicine. However, there is a vast amount of data within even a single individual's genome and finding variants becomes the key challenge in order to pinpoint the best pharmacological treatment for an individual based on their genetic background. Harnessing the combination of data from already existing genomic variant databases and historic clinical data from EMRs, researchers aim to find such cancer-related variations and driver genes. In a few countries, studies revolving around the interaction between the genome and cancer treatment drugs have gained much attention.

In the US, the NCI-MATCH trials, a phase II precision medicine cancer trial initiated in 2015, showed negative results in precision medicine and concluded that the genomic data did not correlate with any significant results in drug variation [113]. This low statistical significance is not surprising from a data mining perspective as numbers of patients accrued for each of the +40 arms within this study were very small, ranging from 4 to 70 people [114]. Furthermore, the majority of the recruited patients (62.5%) had rare tumors that were not the four most common cancers (breast, colorectal, non-small cell lung, and prostate) [115]. This diversity in cancer types may have introduced confounding factors that affected the statistics of the trial.

In Japan, starting 2018, the Japanese Ministry of Welfare and Labor is sponsoring a panel trial on partial genomic testing for oncogenetic variation. This partial genomic testing aims to reveal the best and optimal cancer drug treatment on the individual based on their genetic variations. In 2019, 11 Cancer Genomic Core hospitals and central medical institutions were selected to start collecting genomic data and clinical data in preparation for a nation-wide genomic panel trial [116]. Under the funding of the country's National Health Insurance, it strives to predict cancer patient treatment responses based on their partial genome data.

There is a complex interplay between intricate biological systems and the NCI-MATCH trial illustrates that precision medicine methods need much more development before they can pin point a certain genomic sequences to the onset of cancer. Some have voiced pessimistic views that this precision medicine task is not feasible and overly-costly at this point in time [117]. However, precision medicine is in the horizon. With more data samples, similar research can yield more insight into precision medicine.

In the future, individual whole genome data may be regular practice to include as part of EHRs in order to help deliver the optimal cancer treatment. Currently, there is a bottleneck where there are not enough types of commercialized cancer drug against which to test the genomic variation and to find which treatment works best on an individual. As all aspects of EMR-driven research converge, more medical data will be collected, stored and published. This will lead to already available commercial drugs undergoing more comprehensive pharmacovigilance and real-world data will effectively drive new drug research. Therefore, it is likely that more types of cancer pharmacology products will become available. Furthermore, the efforts in using ML to mine EMRs may lead to AI predicting cancer patient disease trajectories. The trend toward using NLP to extract relevant information from unstructured EMRs and harnessing deep learning could help reproduce drug-related clinical decision making carried out by medical professionals [110, 111].

6.2 Open sourced resources using EMRs in the UK

In England, there are trusts and clinical commissioning groups who oversee how providers such as hospitals and clinics use their resources. A problematic bottlenecks are that different trusts use different EMR platforms, which have little national standardization and do not allow for interprovider access, which especially cause problems when patients switch trust domains.

A remedy to this lack of standardization is the use of open sourced, publicly available resources including de-identified EMR data. Evident from the data-hungry nature of ML methods and their demonstrated need in scalable phenotype-genotype association research, publicly available EMRs play a crucial role in the advancement of this field. Some notable open sourced data sources and tools include the UK Biobank, where 50,000 individuals (aged 40–69) were recruited from England, Wales, Scotland [118]. The biobank includes detailed phenotype and genotype data, lifestyle surveys, pathophysiological data and imaging data on each individual [118]. Once a centralized, open-sourced EMR data is made available, the next step is the development of platforms that interact with said resource.

The CArdiovascular disease research using LInked Bespoke studies and Electronic health Records (CALIBRE) portal offers freely available software that provides tools and algorithms, which is research ready and have already extracted variables extracted from various EMRs. Phenotype algorithms contained in CALIBRE, which employs data from the UK Biobank, are rule based and use phenotype validations like etiological, which use external published evidence to support the algorithm; prognostic, which evaluate the event's similarity to already existing scientific knowledge; case-note review, which compares the positive predictive value (PPV) and the negative predictive value (NPV) against a gold standard like a clinician's notes; cross-EHR-source concordance, which checks the consistency in findings across other EHRs; genetic, which double checks whether there is consistency in genetic associations and external populations, which validates by comparing results to similar studies done in different countries [119]. These phenotype validations, and standardized validation systems in general, are crucial in characterizing ML algorithms since variations in training data can alter outputs even when the ML method does not change. As open source data proliferates, freely available validation methods may grow in a parallel manner.

In addition, openEHR is also a platform that pools industry specifications, clinical models and software that are intended for data science solutions in the healthcare space. OpenEHR was founded in 2003 by an international non-profit organization and maintained by individuals around the world [120]. In 2017, the UK became the first country to introduce infrastructure from openEHR into the main healthcare system to streamline phenotype data collection and vendor-neutral clinical data storage from all the trusts participating in the 100,000 genome project [121]. Newly coordinated pipelines of additional EHR data such as those from the NHS will increase the through-put in openEHR, which in turn develops the best tools to handle big data, which then completes the circle by promoting the use of an ever increasing amount of medical data. This data-driven vision, in which an open community encourages cooperation by open access and pools existing knowledge around EMR-driven healthcare, will certainly accelerate the evolution of ML methods.

6.3 EHR databases in Estonia

Estonia is one of the world-leading countries in terms of the nationwide systematization of digital medical documentation and the high quality of EHRs. By the end

of 2014, Estonia had centralized EHR access via a single portal, where over 99% of the population could view their own medical records [122]. This is a remarkable statistic but more notably, Estonia's EHR vision had already been initiated in 2007 when the Estonian Genome Center of the University of Tartu established the foundations of the Estonian biobank, which includes 52,000 participants worth of genomic and health data representing about 5% of the adult population of Estonia [123, 124]. Seven years later, the Estonian biobank was linked to the Estonian National Health Information System (ENHIS), which included 44,000 inpatient and 212,000 outpatient medical summaries, EHRs and digital prescriptions from all medical service providers [124]. Since the merge, the databases have been updated through periodic additions of EHRs. By 2016, Estonia was ranked within the top three countries to have the best capability of effectively deploying, operating, maintaining and supporting statistical and medical research using EHRs by the HCQI Survey of Electronic Health Record System Development and Use [125]. This extensive data collection was made possible by the national electronic identification card (ID-card) as this chipped ID-card was made compulsory and became part of the national infrastructure [126]. As result of these efforts, Estonian EHR databases are highly valuable sources for researching EHR-driven methods.

An ADE study using Estonian EHR databases by Tasa et al. demonstrates the database's ability to conduct high impact, translational research. The whole-genome sequencing (WGS) data of +2200 Estonian Biobank participants and the EHRs of the sequenced individuals were taken from Health Insurance Fund Treatment Bills, Tartu University Hospital and North Estonia Medical Center databases [127]. EHRs were mined using ICD codes to find ADE occurrences and a mixture of the ICD and manual verification methods was used to identify associations between genetic polymorphisms and ADEs [127]. Associations between genetic variations and drug responses are vital in advancing personalized drug treatment, which is also referred to as pharmacogenomics. Important genes within the study of pharmacogenomics are called pharmacogenes. The study reported 29.1×10^6 novel variants. To priorities genetic analysis, Tasa et al. compiled 1314 loss-of-function, missense, and putative high-impact variants in promoter regions of 64 pharmacogenes [127]. They reported that 80.3% of the variants were rare (MAF < 1%), and this high proportion suggests that gene variation is crucial in understanding pharmacogenomics [127]. Next, the study combined EHRs to the genetic data to extract 1187 participants with potential ADEs. As a validation, Tasa et al. replicated pharmacogenetic associations between the CYP2D6*6 allele and tramadol related ADEs ($p = 0.035$; odds ratio [OR] = 2.67) and between the same allele and amitriptyline induced ADEs ($p = 0.02$; OR = 6.0) [127]. In addition, they replicated four more validated pharmacogenetic associations and discovered nine independent, new gene associations with ADEs in a group of individuals divided by drug prescriptions. Notably, they identified a new association between CTNNA3 and myositis for oxicam-treated participants. This study demonstrated the viability of layering EHR and WGS data at a population-based scale in order to advance pharmacogenomic. Beyond the scope of this study, identifying pharmacogenomic associations relies more and more on big-data driven projects that looks for genetic variants in different communities and highlights variants that can be medically targeted to advance healthcare [128–130].

In summary, Estonia's world-leading efforts to integrate EHRs as a method to feedback data to basic research is a possible future of data-driven healthcare medicine, which focuses on digitization with a vision for translational biomedical research. Estonia created a data-mining driven database, in which different aspects of the EHRs are linked an ID-card. Although different implementations will be necessary to replicate Estonia's rich and accessible EHR database, Estonia sets a

precedent to the rest of the world and demonstrates the positive biomedical implications of such well-organized databases of rich EHR sources.

7. Conclusion

In the past decade, EMRs have become a vital data source in advancing healthcare. In the context of AI, EMRs are highly attractive because there is a vast quantity of rich and variable data types which cannot be processed manually. In the context of biomedical research, EMRs have exciting potential for impactful medical applications, but only if actionable biomedical conclusions can be accurately extracted. In the clinical context, EMRs were introduced to replace the traditional paperwork but were not intended for data-mining research; they were never intended to perform anything that paper documents were not designed to do. Having been introduced in a time before the phrase "machine learning", digitization of medical records has far surpassed the imagined benefits of this transition. Envisioned as a direct replacement of paper records, EMR history has been fraught with difficulties: implementation costs, workflow disruptions and cyber-attacks to name a few. Harnessing EMRs for research purposes marks a milestone in translational biomedical medicine. It is the intersection of basic science, data-driven methods and clinical research where healthcare is transformed: every hospital visit improving human knowledge of diseases one EMR at a time.

The chapter started with a discussion of the EMRs definition, given that they have been introduced with little regard to compatibility with other existing EMR systems. There are many issues that hospitals can encounter when transitioning from paper records to electronic, however, efficiency gains from digitizing records are significant even without the use of big data. To exemplify what can be achieved by applying ML techniques to the data contained in EMRs, three key biomedical research areas were considered: phenotype-genotype association, clinical trials for new drug and pharmacovigilance studies.

Adopting high throughput data strategies into clinical drug trials can reduce the inefficiencies that often plague such trials. EMR mining using already existing systems can improve trial recruitment, but care must be taken to reduce potential bias in patient selection. Additionally, EMRs can be employed to continue data collection after the trial formally ends, a great benefit for financially limited trials, or they can even be treated as a primary data source as long as the data is considered to be of satisfactory standard.

After a drug undergoes clinical trials and is approved for market launch, pharmaceutical companies are encouraged to continue drug surveillance to detect, evaluate and prevent adverse drug events, which create medical and financial burdens. Such surveillance can be cheaply and efficiently done by continually mining EHR narratives. In the context of ADE detection, keyword searches are considered to be too simplistic and to lack scalability. Despite this, they still show some success in small scale studies, serving as a proof of concept that harnessing EHRs with more advanced processes could greatly benefit pharmagovigilance. However, NLP based-approaches performed much better than keyword-based methods and an excellent case study on NLP-driven pharmacovigilance is the MADE1.0 challenge. By bringing together multiple institutions, the challenge succeeded in developing high performing ML methods, including frequent usage of CRFs and LSTM, for the NER and RI tasks. This initiative promoted further works to create even more robust ML methods to extract ADEs from oncology EMRs and reflects the overall trend in the pharmacovigilance space toward CRF, SVM and random forest models.

With this vital context on how ML methods are used to analyze the data within EMRs, some selected international case studies on EHR-driven research were presented. Firstly, on the outlook of oncology precision medicine: NCI-MATCH trials in the US concluded that no drug response is correlated with genomic data, whilst preparation for partial genomic testing for oncology drugs is underway in Japan. Despite negative results nation-wide initiatives may spur on the collective development of drug research. Secondly, UK-based open source resources for EHR manipulation, were discussed, both large consolidated datasets and freely available tools, algorithms and platforms. This vision for open sourced resources is a valuable digital environment in which to pool technical knowledge, especially because of the translational and multi-disciplinary dimension of extracting medically meaningful conclusions from EHRs. Thirdly, the EHR databases set up in Estonia were reviewed, which are both nationally extensive and high quality. This set up the groundwork to deploy a population-based WGS and EHR combinatory study conducive to pharmacogenetic advances. Estonia's databases demonstrate the power of harnessing data from EHR for the progress of healthcare.

In contrast to the recent advancement and current interest in clinically-applied deep learning, there is still no definitive evidence of a model with predictive performance that is similar to a human physician [131]. As of 2020, there is no immediate vision in which AI can fully automate drug research pipelines or independently diagnose and provide subsequent health care procedures making researchers and clinicians obsolete. As we have seen, however, there is ample evidence that EMRs will increasingly play a vital role in all aspects of the drug research arc from fundamental science and clinical trials to post-market surveillance.

Conflict of interest

The author declares no conflict of interest.

Abbreviations

EMR	electronic medical record
EHR	electronic health record
NHS	National Health Services
ML	machine learning
DDI	drug–drug interaction
ADE	adverse drug event
ICD	International Classification of Disease
WHO	World Health Organization
ICD-CM	ICD Clinical Modification
SNP	single nucleotide polymorphism
CNN	convolutional neural network
I/E criteria	inclusion and exclusion criteria
NLP	natural language processing
HMM	hidden Markov model
CRF	conditional random fields
RNN	recurrent neural networks
LSTM	long short-term memory
BiLSTM	bidirectional long short-term memory
NER	named entity recognition

RI	relation identification
SVMs	support vector machines
CALIBRE	CArdiovascular disease research using LInked Bespoke studies and Electronic health Records
WGS	whole-genome sequencing

Author details

Ayaka Shinozaki[1,2,3]

1 techspert.io Ltd, Cambridge, UK

2 Department of Medicine, University of Cambridge, Cambridge, UK

3 Cancer Research UK, Cambridge Institute, Cambridge, UK

*Address all correspondence to: 13shinozakia@gmail.com

References

[1] Sharma H, Mao C, Zhang Y, Vatani H, Liang Y, Zhong Y, et al. Developing a portable natural language processing based phenotyping system. BMC Medical Informatics and Decision Making. 2019;**19**(3):78

[2] Bowton E, Field JR, Wang S, Schildcrout JS, Van Driest SL, Delaney JT, et al. Biobanks and electronic medical records: Enabling cost-effective research. Science Translational Medicine. 2014;**6**(234) 234cm3–234cm3. DOI: 10.1126/scitranslmed.3008604. Available from: https://stm.sciencemag.org/content/6/234/234cm3. ISSN: 1946-6234

[3] Kohane IS. Using electronic health records to drive discovery in disease genomics. Nature Reviews Genetics. 2011;**12**(6):417-428. DOI: 10.1038/nrg2999

[4] Google scholar. Available from: https://scholar.google.com/

[5] Evans RS. Electronic health records: Then, now, and in the future. Yearbook of Medical Informatics. 2016;**25**(S01): S48-S61

[6] Norton PT, Rodriguez HP, Shortell SM, Lewis VA. Organizational influences on health care system adoption and use of advanced health information technology capabilities. The American Journal of Managed Care. 2019;**25**(1):e21

[7] Sachdeva S, Bhalla S. Semantic interoperability in standardized electronic health record databases. Journal of Data and Information Quality (JDIQ). 2012;**3**(1):1-37

[8] Zeng Z, Yu D, Li X, Naumann T, Luo Y. Natural language processing for ehr-based computational phenotyping. IEEE/ACM Transactions on Computational Biology and Bioinformatics. 2018;**16**(1):139-153

[9] Marigorta UM, Rodríguez JA, Gibson G, Navarro A. Replicability and prediction: Lessons and challenges from gwas. Trends in Genetics. 2018;**34**(7): 504-517

[10] Allyn-Feuer A, Higgins GA, Athey BD. Pharmacogenomics in the age of gwas, omics atlases, and phewas. arXiv preprint. arXiv: 1808.09481, 2018

[11] Agarwala V, Khozin S, Singal G, O'Connell C, Kuk D, Li G, et al. Real-world evidence in support of precision medicine: Clinico-genomic cancer data as a case study. Health Affairs. 2018; **37**(5):765-772

[12] Warner JL, Jain SK, Levy MA. Integrating cancer genomic data into electronic health records. Genome Medicine. 2016;**8**(1):113

[13] Qian T, Zhu S, Hoshida Y. Use of big data in drug development for precision medicine: An update. Expert Review of Precision Medicine and Drug Development. 2019;**4**(3):189-200

[14] ICD-11 Implementation or Transition Guide. 2019. Available from: https://icd.who.int/docs/ICD-11ImplementationorTransitionGuide_v105.pdf

[15] Wei W-Q, Bastarache LA, Carroll RJ, Marlo JE, Osterman TJ, Gamazon ER, et al. Evaluating phecodes, clinical classification software, and ICD-9-cm codes for phenome-wide association studies in the electronic health record. PLoS ONE. 2017;**12**(7):1-16. DOI: 10.1371/journal.pone.0175508

[16] Denny JC, Ritchie MD, Basford MA, Pulley JM, Bastarache L, Brown-Gentry

K, et al. PheWAS: Demonstrating the feasibility of a phenome-wide scan to discover gene–disease associations. Bioinformatics. 2010;**26**(9):1205-1210. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btq126

[17] Hebbring SJ, Rastegar-Mojarad M, Ye Z, Mayer J, Jacobson C, Lin S. Application of clinical text data for phenome-wide association studies (PheWASs). Bioinformatics. 2015; **31**(12):1981-1987. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btv076

[18] Liao KP, Sun J, Cai TA, Link N, Hong C, Huang J, et al. High-throughput multimodal automated phenotyping (map) with application to phewas. bioRxiv. 2019. DOI: 10.1101/587436

[19] Yu S, Ma Y, Gronsbell J, Cai T, Ananthakrishnan AN, Gainer VS, et al. Enabling phenotypic big data with phenorm. Journal of the American Medical Informatics Association. 2017; **25**(1):54-60

[20] Coquet J, Bozkurt S, Kan KM, Ferrari MK, Blayney DW, Brooks JD, et al. Comparison of orthogonal NLP methods for clinical phenotyping and assessment of bone scan utilization among prostate cancer patients. Journal of Biomedical Informatics. 2019;**94**: 103184

[21] Zeng Z, Espino S, Roy A, Li X, Khan SA, Clare SE, et al. Using natural language processing and machine learning to identify breast cancer local recurrence. BMC Bioinformatics. 2018; **19**(17):498

[22] Pikoula M, Quint JK, Nissen F, Hemingway H, Smeeth L, Denaxas S. Identifying clinically important COPD sub-types using data-driven approaches in primary care population based electronic health records. BMC Medical Informatics and Decision Making. 2019; **19**(1):86

[23] Zhou S-M, Fernandez-Gutierrez F, Kennedy J, Cooksey R, Atkinson M, Denaxas S, et al. Defining disease phenotypes in primary care electronic health records by a machine learning approach: A case study in identifying rheumatoid arthritis. PLoS ONE. 2016; **11**(5):e0154515

[24] Zeng Z, Deng Y, Li X, Naumann T, Luo Y. Natural language processing for EHR-based computational phenotyping. IEEE/ACM Transactions on Computational Biology and Bioinformatics. 2019. ISSN: 2374-0043; **16**(1):139-153. DOI: 10.1109/TCBB.2018.2849968

[25] Boudellioua I, Kulmanov M, Schofield PN, Gkoutos GV, Hoehndorf R. DeepPVP: Phenotype-based prioritization of causative variants using deep learning. BMC Bioinformatics. 2019;**20**(1):65

[26] Zeng Z, Jiang X, Neapolitan R. Discovering causal interactions using bayesian network scoring and information gain. BMC Bioinformatics. 2016;**17**(1):221

[27] Stark SG, Hyland SL, Fernandes Pradier M, Lehmann K, Wicki A, Perez Cruz F, et al. Unsupervised extraction of phenotypes from cancer clinical notes for association studies. arXiv preprint. arXiv:1904.12973, 2019

[28] Salcedo CC, Labilloy G, Andrew S, Hwa V, Tyzinski L, Grimberg A, et al. OR07–6 integrating targeted bioinformatic searches of the electronic health records and genomic testing identifies a molecular diagnosis in three patients with undiagnosed short stature. Journal of the Endocrine Society. 2019;**3** (Suppl 1). ISSN: 2472-1972. DOI: 10.1210/js.2019-OR07-6

[29] Tong J, Huang J, Chubak J, Wang X, Moore JH, Hubbard RA, et al. An augmented estimation procedure for EHR-based association studies

accounting for differential misclassification. Journal of the American Medical Informatics Association. 2020;**27**(2):244-253. ISSN: 1527-974X. DOI: 10.1093/jamia/ocz180. ocz180

[30] Beaulieu-Jones BK, Greene CS. Semi-supervised learning of the electronic health record for phenotype stratification. Journal of Biomedical Informatics. 2016;**64**:168-178. ISSN: 1532-0464. DOI: 10.1016/j.jbi.2016.10.007

[31] Chiu P-H, Hripcsak G. EHR-based phenotyping: Bulk learning and evaluation. Journal of Biomedical Informatics. 2017;**70**:35-51. ISSN: 1532-0464. DOI: 10.1016/j.jbi.2017.04.009

[32] Hubbard RA, Huang J, Harton J, Oganisian A, Choi G, Utidjian L, et al. A bayesian latent class approach for EHR-based phenotyping. Statistics in Medicine. 2019;**38**(1):74-87. DOI: 10.1002/sim.7953

[33] Beesley LJ, Fritsche LG, Mukherjee B. A modeling framework for exploring sampling and observation process biases in genome and phenome-wide association studies using electronic health records. bioRxiv. 2018. DOI: 10.1101/499392. Available from: https://www.biorxiv.org/content/early/2018/12/20/499392

[34] Meystre S'e M, Heider PM, Kim Y, Aruch DB, Britten CD. Automatic trial eligibility surveillance based on unstructured clinical data. International Journal of Medical Informatics. 2019; **129**:13-19

[35] Hurdle JF, Haroldsen SC, Hammer A, Spigle C, Fraser AM, Mineau GP, et al. Identifying clinical/translational research cohorts: Ascertainment via querying an integrated multi-source database. Journal of the American Medical Informatics Association. 2012;**20**(1):164-171

[36] Obeid JS, Beskow LM, Rape M, Gouripeddi R, Black RA, Cimino JJ, et al. A survey of practices for the use of electronic health records to support research recruitment. Journal of Clinical and Translational Science. 2017;**1**(4): 246-252

[37] Devoe C, Gabbidon H, Schussler N, Cortese L, Caplan E, Gorman C, et al. Use of electronic health records to develop and implement a silent best practice alert notification system for patient recruitment in clinical research: Quality improvement initiative. JMIR Medical Informatics. 2019;7(2):e10020

[38] Bejjanki H, Mramba LK, Beal SG, Radhakrishnan N, Bishnoi R, Shah C, et al. The role of a best practice alert in the electronic medical record in reducing repetitive lab tests. ClinicoEconomics and Outcomes Research: CEOR. 2018;**10**:611

[39] Electronic health records for clinical research (ehr4cr). Available from: http://www.ehr4cr.eu/

[40] Claerhout B, Kalra D, Mueller C, Singh G, Ammour N, Meloni L, et al. Federated electronic health records research technology to support clinical trial protocol optimization: Evidence from ehr4cr and the insite platform. Journal of Biomedical Informatics. 2019; **90**:103090

[41] Liu S, Wang Y, Wen A, Wang L, Hong N, Shen F, et al. Create: Cohort retrieval enhanced by analysis of text from electronic health records using omop common data model. arXiv preprint. arXiv:1901.07601, 2019

[42] CogStack. Cogstack/cogstack-semehr. Available from: https://github.com/CogStack/SemEHR

[43] Glicksberg BS, Miotto R, Johnson KW, Shameer K, Li L, Chen R, et al. Automated disease cohort selection using word embeddings from electronic

health records. In: Pacific Symposium on Biocomputing; World Scientific. 2018. pp. 145-156

[44] Horowitz CR, Sabin T, Ramos M, Richardson LD, Hauser D, Robinson M, et al. Successful recruitment and retention of diverse participants in a genomics clinical trial: A good invitation to a great party. Genetics in Medicine. 2019;**21**: 2364-2370

[45] Devers K, Gray B, Ramos C, Shah A, Blavin F, Waidmann T. The Feasibility of Using Electronic Health Records (EHRs) and Other Electronic Health Data for Research on Small Populations. Urban Institute: Washington; 2013

[46] Chamberlin SR, Bedrick SD, Cohen AM, Wang Y, Wen A, Liu S, et al. Evaluation of patient-level retrieval from electronic health record data for a cohort discovery task. MedRxiv. 2019;**1**:19005280

[47] Aroda VR, Sheehan PR, Vickery EM, Staten MA, LeBlanc ES, Phillips LS, et al. Establishing an electronic health record–supported approach for outreach to and recruitment of persons at high risk of type 2 diabetes in clinical trials: The vitamin D and type 2 diabetes (d2d) study experience. Clinical Trials. 2019; **16**(3):306-315

[48] Pfaff E, Lee A, Bradford R, Pae J, Potter C, Blue P, et al. Recruiting for a pragmatic trial using the electronic health record and patient portal: Successes and lessons learned. Journal of the American Medical Informatics Association. 2018;**26**(1):44-49

[49] Davies G, Jordan S, Brooks CJ, Thayer D, Storey M, Morgan G, et al. Long term extension of a randomised controlled trial of probiotics using electronic health records. Scientific Reports. 2018;**8**(1):7668

[50] Cowie MR, Blomster JI, Curtis LH, Duclaux S, Ford I, Fritz F, et al. Electronic health records to facilitate clinical research. Clinical Research in Cardiology. 2017;**106**(1):1-9

[51] Kibbelaar RE, Oortgiesen BE, Van Der Wal-Oost AM, Boslooper K, Coebergh JW, Veeger NJGM, et al. Bridging the gap between the randomised clinical trial world and the real world by combination of population-based registry and electronic health record data: A case study in haemato-oncology. European Journal of Cancer. 2017;**86**:178-185

[52] Beskow LM, Brelsford KM, Hammack CM. Patient perspectives on use of electronic health records for research recruitment. BMC Medical Research Methodology. 2019;**19**(1):42

[53] Goldstein CE, Weijer C, Brehaut JC, Fergusson DA, Grimshaw JM, Horn AR, et al. Ethical issues in pragmatic randomized controlled trials: A review of the recent literature identifies gaps in ethical argumentation. BMC Medical Ethics. 2018;**19**(1):14

[54] McDermott DS, Kamerer JL, Birk AT. Electronic health records: A literature review of cyber threats and security measures. International Journal of Cyber Research and Education (IJCRE). 2019;**1**(2):42-49

[55] Ganiga R, Pai RM, Pai MMM, Sinha RK. Security framework for cloud based electronic health record (EHR) system. International Journal of Electrical & Computer Engineering. 2020;**10**:2088-8708

[56] Farhadi M, Haddad H, Shahriar H. Compliance checking of open source EHR applications for HIPAA and ONC security and privacy requirements. In: 2019 IEEE 43rd Annual Computer Software and Applications Conference (COMPSAC), Vol. 1; IEEE. 2019. pp. 704-713

[57] Onder G, Pedone C, Landi F, Cesari M, Vedova CD, Bernabei R, et al.

Adverse drug reactions as cause of hospital admissions: Results from the Italian group of pharmacoepidemiology in the elderly (GIFA). Journal of the American Geriatrics Society. 2002; **50**(12):1962-1968. DOI: 10.1046/j.1532-5415.2002.50607.x

[58] Salas-Vega S, Haimann A, Mossialos E. Big data and health care: Challenges and opportunities for coordinated policy development in the eu. Health Systems & Reform. 2015; **1**(4):285-300

[59] Mehta N, Pandit A. Concurrence of big data analytics and healthcare: A systematic review. International Journal of Medical Informatics. 2018;**114**:57-65

[60] Harpaz R, Callahan A, Tamang Y, Low S, Odgers D, Finlayson S, et al. Text mining for adverse drug events: The promise, challenges, and state of the art. Drug Safety. 2014. ISSN: 1179-1942; **37**(10):777-790. DOI: 10.1007/s40264-014-0218-z

[61] Cami A, Arnold A, Manzi S, Reis B. Predicting adverse drug events using pharmacological network models. Science Translational Medicine. 2011. ISSN: 1946-6234;**3**(114) 114ra127–114ra127. DOI: 10.1126/scitranslmed.3002774. Available from: https://stm.sciencemag.org/content/3/114/114ra127

[62] Pouliot Y, Chiang AP, Butte AJ. Predicting adverse drug reactions using publicly available pubchem bioassay data. Clinical Pharmacology & Therapeutics. 2011;**90**(1):90-99. DOI: 10.1038/clpt.2011.81

[63] Liu M, Wu Y, Chen Y, Sun J, Zhao Z, Chen X-w, et al. Large-scale prediction of adverse drug reactions using chemical, biological, and phenotypic properties of drugs. Journal of the American Medical Informatics Association. 2012;**19**(e1): e28-e35

[64] Zheng H, Wang H, Xu H, Wu Y, Zhao Z, Azuaje F. Linking biochemical pathways and networks to adverse drug reactions. IEEE Transactions on Nanobioscience. June 2014. ISSN: 1558-2639;**13**(2):131-137. DOI: 10.1109/TNB.2014.2319158

[65] Harpaz R, Vilar S, DuMouchel W, Salmasian H, Haerian K, Shah NH, et al. Combing signals from spontaneous reports and electronic health records for detection of adverse drug reactions. Journal of the American Medical Informatics Association. 2012;**20**(3): 413-419

[66] Nadkarni PM. Drug safety surveillance using de-identified EMR and claims data: Issues and challenges. Journal of the American Medical Informatics Association. 2010. ISSN: 1527-974X (Electronic); 1067–5027 (Print); 1067–5027 (Linking);**17**(6): 671-674. DOI: 10.1136/jamia.2010.008607

[67] Classen DC, Resar R, Griffin F, Federico F, Frankel T, Kimmel N, et al. 'Global trigger tool' shows that adverse events in hospitals may be ten times greater than previously measured. Health Affairs. 2011;**30**(4):581-589. DOI: 10.1377/hlthaff.2011.0190

[68] Doupi P. Using EHR data for monitoring and promoting patient safety: Reviewing the evidence on trigger tools. Studies in Health Technology and Informatics. 2012;**180**: 786-790

[69] Luo Y, Riedlinger G, Szolovits P. Text mining in cancer gene and pathway prioritization. Cancer Informatics. 2014;**13**(Suppl 1):69-79

[70] Cohen KB, Demner-Fushman D. Biomedical Natural Language Processing. Amsterdam, The Netherlands: John Benjamins; 2014. Available from: https://www.jbe-platform.com/content/books/9789027271068

[71] Luo Y, Thompson WK, Herr TM, Zeng Z, Berendsen MA, Jonnalagadda SR, et al. Natural language processing for EHR-based pharmacovigilance: A structured review. Drug Safety. 2017;**40**(11): 1075-1089. ISSN: 1179-1942. DOI: 10.1007/s40264-017-0558-6

[72] Leeper NJ, Bauer-Mehren A, Iyer SV, LePendu P, Olson C, Shah NH. Practice-based evidence: profiling the safety of cilostazol by text-mining of clinical notes. PLoS ONE. 2013;**8**(5): e63499

[73] Banda JM, Callahan A, Winnenburg R, Strasberg HR, Cami A, Reis BY, et al. Feasibility of prioritizing drug–drug-event associations found in electronic health records. Drug Safety. 2016;**39**(1):45-57

[74] Ferrajolo C, Verhamme KMC, Trifirò G, Jong G W't, Giaquinto C, Picelli G, et al. Idiopathic acute liver injury in paediatric outpatients: Incidence and signal detection in two European countries. Drug Safety. 2013; **36**(10):1007-1016

[75] Ferrajolo C, Coloma PM, Verhamme KMC, Schuemie MJ, de Bie S, Gini R, et al. Signal detection of potentially drug-induced acute liver injury in children using a multi-country healthcare database network. Drug Safety. 2014;**37**(2):99-108

[76] Pathak J, Kiefer RC, Chute CG. Using linked data for mining drug-drug interactions in electronic health records. Studies in Health Technology and Informatics. 2013;**192**:682

[77] Pathak J, Kiefer RC, Chute CG. Mining drug-drug interaction patterns from linked data: A case study for warfarin, clopidogrel, and simvastatin. In: 2013 IEEE International Conference on Bioinformatics and Biomedicine; IEEE. 2013. pp. 23-30

[78] Honigman B, Lee J, Rothschild J, Light P, Pulling RM, Yu T, et al. Using computerized data to identify adverse drug events in outpatients. Journal of the American Medical Informatics Association. 2001;**8**(3):254-266

[79] Honigman B, Light P, Pulling RM, Bates DW. A computerized method for identifying incidents associated with adverse drug events in outpatients. International Journal of Medical Informatics. 2001. ISSN: 1386-5056; **61**(1):21-32. DOI: 10.1016/S1386-5056 (00)00131-3. Available from: http:// www.sciencedirect.com/science/article/ pii/S1386505600001313

[80] Chazard E, Băceanu A, Ferret L, Ficheur G. The ADE scorecards: A tool for adverse drug event detection in electronic health records. Studies in Health Technology and Informatics. 2011;**166**:169-179

[81] Chazard E, Ficheur G, Bernonville S, Luyckx M, Beuscart R. Data mining to generate adverse drug events detection rules. IEEE Transactions on Information Technology in Biomedicine. 2011;**15**(6): 823-830

[82] Epstein RH, Jacques PS, Stockin M, Rothman B, Ehrenfeld JM, Denny JC. Automated identification of drug and food allergies entered using non-standard terminology. Journal of the American Medical Informatics Association. 2013;**20**(5):962-968

[83] Eriksson R, Jensen PB, Frankild S, Jensen LJ, Brunak S. Dictionary construction and identification of possible adverse drug events in Danish clinical narrative text. Journal of the American Medical Informatics Association. 2013;**20**(5):947-953

[84] Eriksson R, Werge T, Jensen LJ, Brunak S. Dose-specific adverse drug reaction identification in electronic patient records: Temporal data mining

in an inpatient psychiatric population. Drug Safety. 2014;**37**(4):237-247

[85] Roitmann E, Eriksson R, Brunak S. Patient stratification and identification of adverse event correlations in the space of 1190 drug related adverse events. Frontiers in Physiology. 2014;**5**:332

[86] Friedman C, Shagina L, Lussier Y, Hripcsak G. Automated encoding of clinical documents based on natural language processing. Journal of the American Medical Informatics Association. 2004;**11**(5):392-402

[87] Aronson AR. Effective mapping of biomedical text to the umls metathesaurus: The metamap program. In: Proceedings of the AMIA Symposium. American Medical Informatics Association. 2001. p. 17

[88] Savova GK, Masanz JJ, Ogren PV, Zheng J, Sohn S, Kipper-Schuler KC, et al. Mayo clinical text analysis and knowledge extraction system (ctakes): Architecture, component evaluation and applications. Journal of the American Medical Informatics Association. 2010; **17**(5):507-513

[89] Re'ategui R, Ratt'e S. Comparison of metamap and ctakes for entity extraction in clinical notes. BMC Medical Informatics and Decision Making. 2018;**18**(3):74

[90] Xu H, Stenner SP, Doan S, Johnson KB, Waitman LR, Denny JC. Medex: A medication information extraction system for clinical narratives. Journal of the American Medical Informatics Association. 2010;**17**(1): 19-24

[91] Cunningham H. Gate, a general architecture for text engineering. Computers and the Humanities. 2002; **36**(2):223-254

[92] Ritwik B, Ramakrishnan IV, Henry M, Perciavalle M. Patient centered identification, attribution, and ranking of adverse drug events. In: 2015 International Conference on Healthcare Informatics. IEEE. 2015. pp. 18-27

[93] Liu Y, LePendu P, Iyer S, Shah NH. Using temporal patterns in medical records to discern adverse drug events from indications. AMIA Summits on Translational Science Proceedings. 2012; **47**:2012

[94] LePendu P, Iyer SV, Bauer-Mehren A, Harpaz R, Mortensen JM, Podchiyska T, et al. Pharmacovigilance using clinical notes. Clinical Pharmacology & Therapeutics. 2013;**93**(6):547-555

[95] Sun W, Rumshisky A, Uzuner O. Temporal reasoning over clinical text: The state of the art. Journal of the American Medical Informatics Association. 2013;**20**(5):814-819

[96] Umass bionlp projects. Available from: https://bio-nlp.org/index.php/projects/39-nlpchallenges

[97] Jagannatha A, Liu F, Liu W, Yu H. Overview of the first natural language processing challenge for extracting medication, indication, and adverse drug events from electronic health record notes (made 1.0). Drug Safety. 2019;**42**(1):99-111

[98] Sutton C, McCallum A, et al. An introduction to conditional random fields. Foundations and Trends® in Machine Learning. 2012;**4**(4):267-373

[99] Olah C. Understanding LSTM networks. Available from: https://colah.github.io/posts/2015-08-Understanding-LSTMs/

[100] Wunnava S, Qin X, Kakar T, Rundensteiner EA, Kong X. Bidirectional LSTM-CRF for adverse drug event tagging in electronic health records. In: International Workshop on Medication and Adverse Drug Event Detection. 2018. pp. 48-56

[101] Berwick R. An idiot's guide to support vector machines (SVMS). Available from: http://web.mit.edu/6.034/wwwbob/

[102] Chapman AB, Peterson KS, Alba PR, DuVall SL, Patterson OV. Detecting adverse drug events with rapidly trained classification models. Drug Safety. 2019;**42**(1):147-156

[103] Burgersmoke. burgersmoke/made-crf. 2019

[104] Chapman AB, Peterson KS, Alba PR, DuVall SL, Patterson OV. Hybrid system for adverse drug event detection. In: International Workshop on Medication and Adverse Drug Event Detection. 2018. pp. 16-24

[105] Dandala B, Joopudi V, Devarakonda M. Adverse drug events detection in clinical notes by jointly modeling entities and relations using neural networks. Drug Safety. 2019; **42**(1):135-146

[106] Li F, Liu W, Hong Y. Extraction of information related to adverse drug events from electronic health record notes: Design of an end-to-end model based on deep learning. JMIR Medical Informatics. 2018;**6**(4):e12159

[107] Aramaki E, Miura Y, Tonoike M, Ohkuma T, Masuichi H, Waki K, et al. Extraction of adverse drug effects from clinical records. Medinfo. 2010;**160**: 739-743

[108] Henriksson A, Zhao J, Boström H, Dalianis H. Modeling electronic health records in ensembles of semantic spaces for adverse drug event detection. In: 2015 IEEE International Conference on Bioinformatics and Biomedicine (BIBM); IEEE. 2015. pp. 343-350

[109] Wang G, Jung K, Winnenburg R, Shah NH. A method for systematic discovery of adverse drug events from clinical notes. Journal of the American Medical Informatics Association. 2015; **22**(6):1196-1204

[110] Srivastava S, Soman S, Rai A, Srivastava PK. Deep learning for health informatics: Recent trends and future directions. In: 2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI); IEEE. 2017. pp. 1665-1670

[111] Wu Y, Warner JL, Wang L, Jiang M, Xu J, Chen Q, et al. Discovery of noncancer drug effects on survival in electronic health records of patients with cancer: A new paradigm for drug repurposing. JCO Clinical Cancer Informatics. 2019;**3**:1-9

[112] Chang Y, Park H, Yang H-J, Lee S, Lee K-Y, Kim TS, et al. Cancer drug response profile scan (CDRscan): A deep learning model that predicts drug effectiveness from cancer genomic signature. Scientific Reports. 2018;**8**(1): 8857

[113] Nci-match precision medicine clinical trial. Available from: https://www.cancer.gov/about-cancer/treatment/clinical-trials/nci-supported/nci-match

[114] Nci-match/eay131-ecog-acrin. 2020. Available from: https://ecog-acrin.org/trials/nci-match-eay131

[115] Nci-match trial releases new findings. Available from: https://www.cancer.gov/news-events/press-releases/2018/nci-match-first-results

[116] Oncoguide NCC oncopanel system insurance developed by the national cancer center. Available from: https://www.ncc.go.jp/jp/information/pr_release/2019/0529/index.html

[117] OF PRECISION. The Precision-Oncology Illusion. 2016

[118] Uk biobank. Available from: https://www.ukbiobank.ac.uk/

[119] Denaxas S, Parkinson H, Fitzpatrick N, Sudlow C, Hemingway H. Analyzing the heterogeneity of rule-based EHR phenotyping algorithms in caliber and the Uk biobank. BioRxiv. 2019:685156

[120] Open industry specifications, models and software for e-health

[121] Heard S, Beale T. Available from: https://www.openehr.org/openehr_in_use/deployed_solutions_detail/27

[122] Gheorghiu B, Hagens S. Use and maturity of electronic patient portals. Studies in Health Technology and Informatics. 2017:136-141

[123] Leitsalu L, Haller T, Esko T, Tammesoo M-L, Alavere H, Snieder H, et al. Cohort profile: Estonian biobank of the Estonian Genome Center, University of Tartu. International Journal of Epidemiology. 2014;**44**(4): 1137-1147

[124] Leitsalu L, Alavere H, Tammesoo M-L, Leego E, Metspalu A. Linking a population biobank with national health registries—The estonian experience. Journal of Personalized Medicine. 2015; **5**(2):96-106

[125] Oderkirk J. Readiness of Electronic Health Record Systems to Contribute to National Health Information and Research. 2017

[126] Sepper R, Ross P, Tiik M. Nationwide health data management system: A novel approach for integrating biomarker measurements with comprehensive health records in large populations studies. Journal of Proteome Research. 2010;**10**(1):97-100

[127] Tasa T, Krebs K, Kals M, Mägi R, Lauschke VM, Haller T, et al. Genetic variation in the estonian population: Pharmacogenomics study of adverse drug effects using electronic health records. European Journal of Human Genetics. 2019;**27**(3):442

[128] Esplin ED, Oei L, Snyder MP. Personalized sequencing and the future of medicine: Discovery, diagnosis and defeat of disease. Pharmacogenomics. 2014;**15**(14):1771-1790

[129] Relling MV, Evans WE. Pharmacogenomics in the clinic. Nature. 2015;**526**(7573):343

[130] Ramos E, Doumatey A, Elkahloun AG, Shriner D, Huang H, Chen G, et al. Pharmacogenomics, ancestry and clinical decision making for global populations. The Pharmacogenomics Journal. 2014; **14**(3):217

[131] Chen D, Liu S, Kingsbury P, Sohn S, Storlie CB, Habermann EB, et al. Deep learning and alternative learning strategies for retrospective real-world clinical data. NPJ Digital Medicine. 2019;**2**(1):1-5