HHS Public Access

Author manuscript

Drug Discov Today. Author manuscript; available in PMC 2021 November 27.

Published in final edited form as:

Drug Discov Today. 2021 May ; 26(5): 1256–1264. doi:10.1016/j.drudis.2020.12.013.

Applications of artificial intelligence in drug development using real-world data

Zhaoyi Chen[1,‡], **Xiong Liu**[2,‡], **William Hogan**[1], **Elizabeth Shenkman**[1], **Jiang Bian**[1]

[1]Department of Health Outcomes and Biomedical Informatics, College of Medicine, University of Florida, Gainesville, FL 32610-0177, USA

[2]AI Innovation Center, Novartis, Cambridge, MA 02142, USA

Abstract

The US Food and Drug Administration (FDA) has been actively promoting the use of real-world data (RWD) in drug development. RWD can generate important real-world evidence reflecting the real-world clinical environment where the treatments are used. Meanwhile, artificial intelligence (AI), especially machine- and deep-learning (ML/DL) methods, have been increasingly used across many stages of the drug development process. Advancements in AI have also provided new strategies to analyze large, multidimensional RWD. Thus, we conducted a rapid review of articles from the past 20 years, to provide an overview of the drug development studies that use both AI and RWD. We found that the most popular applications were adverse event detection, trial recruitment, and drug repurposing. Here, we also discuss current research gaps and future opportunities.

Introduction

Drug development is the process of bringing a new drug molecule into clinical practice; in its broadest definition, it includes all stages from the basic research of finding a suitable molecular target to large-scale Phase III clinical studies that support the commercial launch of the drug to post-market pharmacosurveillance and drug-repurposing studies [1,2]. During the drug development process, chemical entities that have the potential to become therapeutic agents are identified and thoroughly tested, and the entire process is lengthy and costly. It is estimated that, for every new drug brought to the market, it typically costs billions of US dollars and >10 years of work [3,4]. Therefore, strategies that can facilitate and accelerate the drug development process are of high interest.

Recently, the FDA has been actively promoting the use of RWD for drug development [5,6]. The term ‘RWD’ refers to data collected from sources outside of conventional

Corresponding author: Bian, J. (bianjiang@ufl.edu).

‡These authors contributed equally to this work and are co-first authors.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at https://doi.org/10.1016/j.drudis.2020.12.013.

research settings, including electronic health records (EHRs), administrative claims, and billing data, among others [5–7]. These RWD often contain detailed patient information, such as disease status, treatment, treatment adherence and outcomes, comorbidities, and concurrent treatments that are tracked longitudinally. The information generated from RWD can provide important real-world evidence to inform therapeutic development, outcomes research, patient care, safety surveillance, and comparative effectiveness studies [8]. More importantly, the use of RWD allows clinical researchers and regulatory agencies to answer questions more efficiently, saving time and money while yielding answers that are generalizable to the broader population. Over the past decade, there has been an increased uptake of EHR systems in the USA. These technological advances and policy changes in the USA have created a fertile ground with increasing opportunities to use RWD to facilitate drug development. Thus, the FDA has provided guidance on the use of EHR data in clinical investigations [5] as well as guidance on incorporating RWD into regulatory submissions to the FDA [9].

By contrast, the field of AI, including ML/DL, has moved from largely theoretical studies to real-world applications thanks to both the exponential growth of computing power and advances in AI methods [10]. AI has been widely used in many stages of the drug development process to identify novel targets [11], increase understanding of disease mechanisms [12], and develop new biomarkers [13], among others. Many pharmaceutical companies have begun to invest in resources, technologies, and services, especially in generating and assembling data sets to support research in AI and ML/DL, and many of those data sets are from RWD sources. There is an emerging need for an overview of the intersection between AI and RWD in current drug development studies to describe the current trends, identify existing research gaps, and provide insights into potential future directions. Thus, we conducted a rapid review summarizing published articles related to the intersection of AI, RWD, and drug development over the past 20 years. Our specific aims were to identify current trends in using AI and RWD in drug development studies and, subsequently, any challenges and opportunities?

Literature search

Definations of drug development, AI, and RWD

The drug development process, according to the FDA's definition [14], has four stages: (i) drug discovery: the discovery of new therapeutic agents through the understanding of disease mechanisms and properties of molecular compounds (or other technologies); (ii) preclinical research: laboratory and animal testing to answer questions about the safety of the new drug targets; (iii) clinical research: different stages of clinical trials to test the new drug on humans to assess its safety and efficacy; and (iv) postmarketing research: pharmacosurveillance and comparative effectiveness studies.

The definition of AI methods is less clear and varies in computer science and informatics literature. In this rapid review, we chose the definition 'the use of complex algorithms and software to emulate human cognition in the analysis of complicated medical data, and analyse the relationships between prevention or treatment techniques and patient outcomes.' [15] To be more concrete, the specific AI-related methods we considered include ML and

DL (a subbranch of ML), which are in general accepted by different research communities as AI tasks [16].

In terms of RWD, the FDA defines RWD as 'the data relating to patient health status and/or the delivery of health care routinely collected from a variety of sources', which include patient EHRs and claims data, as well as other patient-generated health data, such as those generated in home-use care settings and data from mobile devices that can inform health status [7,8]. Here, because we aim to understand RWD that can be used to support drug development, we focus on RWD sources that provide clinical data not collected in interventional, controlled, experimental clinical research settings [e.g., randomized controlled trials (RCTs)], which include data generated not only from the delivery of routine care (e.g., EHR, claims databases, or disease registries) but also from study designs that can generate RWD (e.g., observational studies and pragmatic clinical trials) [17]. We exclude RWD that are generated from personal devices, such as smartphones and activity trackers.

Eligibility criteria

The inclusion criteria for our review were: (i) studies using RWD as data sources; (ii) studies using AI methods for statistical analysis or data mining; and (iii) studies focused on the development of drugs. As a rapid review, we first focused on identifying existing review articles.

Search strategy and study selection

We performed a literature search through PubMed to identify relevant review articles published until July 1, 2020. In our search strategy, we considered different combinations of search keywords dictated by the definitions of RWD, AI, and drug development that we chose to focus on. Our search query included three distinct sets of keywords for RWD, AI, and different stages of the drug development process, respectively. For completeness, we included keywords such as 'natural language processing' in the AI keywords, because state-of-the-art models for these NLP tasks are often ML/DL methods. The full search query and the complete list of keywords are in Table S1 in the supplemental information online.

Following best practice for rapid reviews [18,19], we first restricted our search to identify existing review articles for inclusion. We then manually identified the specific AI and RWD applications described in these reviews. Next, based on the identified applications, we performed a second round of literature search to look for their detailed approaches, including data source, data type, and analytical methods used. Figure 1 summarizes the overall search and screening process.

Current progress in the literature

In the first round of literature search, a total of 23 review articles were identified; among them, 16 met our inclusion criteria. Based on these review papers, we first highlight the key steps in the drug development process and then summarize the identified research topics in each step (Fig. 2a). We then summarize the applications that used RWD (Fig. 2b) and AI + RWD (Fig. 2c) to address these research questions.

Drug development process and applications of real-world data

The first step in the drug development process is the discovery of potential therapeutic agents, where researchers investigate the interactions among different molecules, genes, and proteins, and then identify which molecules have high potential with the goal of finding novel targets, biomarkers, and compounds [14]. Some of these goals can be achieved using RWD applications. For example, in a recent review paper [17], Singh *et al.* identified 20 studies that used RWD to facilitate drug discovery and clinical research. Among them, 16 identified or validated new phenotypes, disease markers, and biomarkers for patient identification and stratification.

The next step is preclinical testing, which includes both *in vitro* and *in vivo* testing. In this stage, the safety of drug molecules is tested in test tubes, living cell cultures, and animal models. This is a crucial step because the drug development can only move into human trials with extensive data on safety in preclinical research. In the review papers we included, there were no studies identified for this stage.

After the preclinical testing, once the Investigational New Drug (IND) application is approved, drug development moves into clinical research stages. There are three phases of clinical studies before the drug can be submitted for marketing approval. The key issue that needs to be addressed in this step is to evaluate both the safety and efficacy of the new agents in the target human population [20]. RCTs are still the gold standard to generate clinical evidence; however, RWD have become an important data source for RCTs to understand how the developed treatments are being used in real-world settings. For example, Lai *et al.* examined the impact of using EHRs for clinical research recruitment in a review of 13 research articles [21]. They found that the automation in screening and patient identification could contribute to higher recruitment yield and reduced workload.

After a drug is available on the market, the drug developers are required to submit regular reports detailing adverse events (AEs) associated with the drug [14]. In addition to AE reporting, observational studies and pragmatic clinical trials are also conducted using RWD to evaluate the safety of the drug in real-world settings. For pharmacosurveillance, RWD has gained significant attention in recent years. For example, in 2012, Warrer *et al.* conducted a review on studies that used text-mining techniques on narrative documents to investigate AEs [22], where only seven studies were identified. In a more recent review by Luo *et al.* in 2017 on the same topic, 48 studies were identified [23]. These studies showed that text-mining techniques, ranging from simple free-text searching to more advanced ML/DL-based natural language processing (NLP) methods, can be powerful in AE detection, given that AEs are more extensively documented in EHR narratives.

Applications of AI methods using RWD in the drug development process

Across the different drug development stages, few studies used AI on RWD, and most were found in the clinical or postmarketing stage. Three main types of study used AI on RWD (Fig. 2c): trial recruitment optimization, AE detection, and drug repurposing. Therefore, we conducted a second literature search focusing on individual research studies of these three main applications (Fig. 1). Similar to the first literature search, we screened all studies on

these three topics using keywords related to AI and RWD, as detailed earlier. A total of 65 research studies were included after title/abstract and full-text screening. In Table 1, we summarize these studies into subcategories with examples [24–28]. In Figure 3, we show the increasing trend of studies that use AI methods with RWD in the drug development process over the past 15 years. Overall, we observed a steady increase in the total number of studies. In particular, the number of studies focusing on AE detection has exploded and many focused on using NLP methods to extract AE from free-text narratives, likely because of advances in DL-based NLP methods that achieved state-of-the-art performance [29]. Nevertheless, we also observed more studies that tried to leverage AI methods on RWD for optimizing clinical trial recruitments. Moreover, clinical drug repurposing has emerged as a new application area in the drug development process.

Figure 4 summarizes the numbers and percentages of different data sources, data types, and AI methods being used in the 65 studies. Given the overwhelming number of studies used AI-driven NLP methods, we separated NLP studies from other ML/DL studies. State-of-the-art NLP methods often leverage ML and DL approaches such as BERT [29,30]. Overall, EHR data were the most popular data source, especially unstructured clinical notes. Consequently, a large number of studies have focused on developing or using NLP methods. Among the 55 studies on AE detection, 41 (74.5%) were NLP related. Some studies developed a NLP system to extract information from clinical notes to identify AEs related to the administration of medication. For example, Yang *et al.* developed a Long Short-Term Memory (LSTM)-based DL model to detect medication, AEs, and their relations from clinical text [31]. In other studies, the AEs and associated attributes (e.g., severity) extracted from the NLP pipeline were further fed into a downstream model to assess association between AEs and other health outcomes. For example, Zhang *et al.* first used NLP to identify patients who had AEs related to statin therapy, and then examined the relationship between continuation of statin therapy and incidence of death and cardiovascular events among these patients [32]. Meanwhile, most studies on recruitment optimization (75% of studies included) also utilized clinical notes from EHR data, and attempted to identify eligible populations for trials using information extracted from NLP. For example, Spasic *et al.* used an NLP system that combined rule-based knowledge infusion and ML algorithms to analyze longitudinal patient records to determine whether the corresponding patients met given eligibility criteria for clinical trials [33]. Finally, of the two articles on clinical drug repurposing [34,35], one used NLP methods. In work by Xu *et al.*, automated informatics methods, including NLP, were used on EHR data to identify patient cohorts and medication information [34]; the authors then assessed whether metformin is a potential drug that can be repurposed to cancer treatment. In the other clinical drug-repurposing study, Kuang *et al.* developed a ML-based drug repurposing approach, called baseline regularization, to predict the effects of drugs on different physical measurements, such as fasting blood glucose [35], to identify potential repurposing. Although there is a wealth of literature on drug repurposing using EHRs, few studies have used advanced AI methods, with most using traditional statistical approaches, such as Cox regression [36].

Current trends in AI methods on RWD in drug development research

We identified 16 review articles related to the use of AI methods on RWD published over the past 20 years and an increasing number of original studies in three main application areas: AE detection, recruitment optimization, and drug repurposing.

The most common application area that used AI on RWD was for AE detection, primarily focusing on using NLP on unstructured clinical notes from EHR. The reasons for such a rising popularity are twofold: (i) the abundance of textual information in RWD, especially EHRs; and (ii) the rapid advancement in NLP methods, especially those new DL-based models with state-of-the-art performance. In fact, >80% of the clinical information in EHR is documented in free-text [37], which makes text mining an ideal tool. EHRs have been particularly useful for investigating AEs and other therapeutic effects because of their continuous and longitudinal nature of clinically relevant outcomes and medication exposures.

We also identified several studies that focused on recruitment optimization and drug repurposing. These tasks are suitable for the use of AI and RWD because: (i) the extensive collections of RWD provide sufficient sample sizes to identify individuals who meet recruitment criteria; (ii) the longitudinal detailed medical histories of patients captured in these RWD sources make it possible for researchers to identify drugs that might be effective for indications other than the primary use; (iii) AI and data-driven approaches could potentially minimize the selection bias because they do not rely on researchers' predetermined assumptions, and, thus, are able to identify novel associations that were previously unknown; and (iv) modern AI methods are capable of handling the high dimensionality and complexity of RWD as well as the complex combinations and interactions of RWD variables.

Challenges and future directions

Challenges of using AI and RWD in the drug development studies

First, one major challenge is the quality of the data in many RWD sources. For example, information heterogeneity has been reported in EHRs because clinicians do not always document the care in the same way [38]. Such variance makes it difficult to extract the same information (e.g., outcome measures) consistently. Other data-inconsistency issues, such as missing data and selection bias, also present significant challenges to researchers because data collection in real-world settings is usually heterogeneous and unstandardized. Second, most of the studies we identified focused on prediction or classification tasks and often overemphasized model performance rather than learning the casual effects [39,40]. Furthermore, most of these existing studies do not integrate *a priori* causal knowledge to guide the learning process and, as a result, no causal relationship can be estimated. Third, the transportability and interpretability of these studies also need to be further assessed. External validations using independent sources to ensure the findings are representative and generalizable are recommended, but such validation studies are often difficult to execute for multiple reasons, including: (i) sharing of individual-level clinical data remains difficult because of not only ethical and legal issues, but also market competition concerns; and

(ii) the lack of standardization and harmonization across the different data sources (e.g., inconsistent outcome measures), making replication studies unattainable.

Nevertheless, significant advancements have also been made to tackle these challenges. First, advances in AI methods, especially in DL, have prompted studies that consider heterogeneous data sources and types (e.g., clinical data, imaging, -omics data, and knowledge bases, among others) in one coherent model. Li *et al.* developed a DL model based on recurrent neural networks to learn representation and temporal dynamics of longitudinal cognitive measures of individual subjects and combined them with baseline hippocampal magnetic resonance imaging (MRI) measures to build a prognostic model of Alzheimer's disease dementia progression [41]. Other developments in DL include the ability to handle not only the temporal order of clinical events, but also the long-term dependencies among the events as well as the time-varying effects of the covariates. For example, time-aware LSTM (T-LSTM) incorporates elapsed time information into the standard LSTM architecture to handle irregular time intervals in longitudinal EHR data [42] to learn disease subphenotypes. BEHRT, a new deep neural sequence transduction model for prediction of interpretable personalized risk using EHR data, models the temporal evolution of EHR data through utilizing various forms of sequential concept and enabled the incorporation of multiple heterogeneous concepts (e.g., diagnosis, medication, measurements, and more) to further improve the accuracy of its predictions [43]. In NLP, new methods have been developed that can incorporate factual medical knowledge from existing ontologies/knowledge bases (e.g., the Unified Medical Language System) to further improve the performance of NLP tasks, such as for clinical concept extraction [44].

Second, the use of causal modeling tools in AI, such as causal diagrams, could provide important additions to the implementations of causal inference using RWD. Causal modeling can also lead to improvements in the interpretability and adaptability of AI models in these drug development studies [45]. This concept of causal AI has been applied successfully in public health studies, such as the identification of occupational risk factors [46,47] and the prediction of diarrhea incidence in children [48], among others, and could be used in future drug development research, such as the 'target trial' [49] framework aiming to establish causal treatment effects using RWD without conducting RCTs. Additionally, the emerging of explainable AI (XAI) could help to interpret and understand AI decisions. The XAI models use different mechanisms (e.g., feature interaction and importance, knowledge distillation, and rule extraction) on top of ML/DL models to generate interpretable outputs, such as variable ranking [50], which ultimately help us understand why an AI system makes a certain decision. XAI models are particular useful for tasks such as drug repurposing because these tasks are generating hypotheses for which plausible explanations are crucial.

Finally, the establishment of large research networks, such as the national Patient-Centered Clinical Research Network (PCORnet) [51], Observational Health Data Sciences and Informatics (OHDSI) consortium [52], and the Clinical and Translational Service Award Accrual to Clinical Trails (CTSA ACT) network [53], facilitate the sharing of RWD. Each of these large networks comprises multiple sites across the USA and internationally, and the same data infrastructure (i.e., the same ontologies and common data models) are being used in each network. RWD from these networks represent a diverse set of patients and

institutions and provide the opportunities to conduct large populational studies to understand factors that contribute to health and illness in a heterogeneous and real-world setting. In addition, de-identification strategies, such as those for automated de-identification of massive clinical notes [54], have been widely applied to facilitate data sharing across different institutions. Furthermore, privacy-preserving record-linkage tools have showed high precisions in linking and deduplicating patient records without sharing of protected identifiable information [55]. Although these de-identification strategies might not be applicable for every data type, they provide capabilities to facilitate data sharing across sites and integration of different data sources.

Future applications

There are several other scenarios where RWD and AI methods might be useful in the drug development process. For example, traditionally, clinical trial simulation (CTS) studies use computerized simulation methods on virtual populations to test different trial designs before resources are invested in conducting the actual clinical trial [56]. CTS that incorporates RWD can simulate its virtual populations more realistically. Furthermore, recent developments in the 'target trial' framework, emulating hypothetical trials with RWD, enable us to identify unbiased initiation of exposures and reach an unbiased estimation of the casual relationships [49]. Combing the concept of modern trial emulation and the traditional CTS approaches, a trial simulation framework with RWD that can systematically test the different assumptions of a clinical trial to inform future trial design and produce causal results from RWD will be of high interest.

To facilitate the discovery of new drug targets, another emerging trend is the linkage of EHRs with other data sources, such as biobanking data, to study drug–phenotype and drug– gene interactions. For example, researchers from the Vanderbilt Electronic Systems for Pharmacogenomic Assessment (VESPA) Project [57], demonstrated that EHR-based biobanks could be cost-effective tools for establishing disease and drug associations, because such applications allow the reuse of biological samples for multiple studies without incremental collection, extraction, or processing costs, and the integration with EHR system allows for centralized de-identification and phenotype annotations.

Finally, we highlight the importance of the clinical and translational science life cycle in the drug development process. For example, the drug-repurposing signals identified from population-based studies will need to be looped back to the preclinical and clinical study stages for further validation and evaluation [58].

Limitations of our work

First, as a rapid review, our work is not comprehensive, but has provided a rapid and necessary summary and discussion of the topic. Second, our definition of AI is restricted to ML/DL methods (and their applications in NLP), and our definition of RWD is constrained to clinical data generated from the delivery of routine care (e.g., EHRs and claims data). Therefore, studies using AI methods such as automation and studies using data from personal devices, such as social media and activity trackers, were not included in our review. For example, social media data have shown promise in identifying AEs, although the noisy

nature of social media data remains as a challenge [59,60]. These computational methods and data sources could provide additional insights into the drug development process and should be revisited in a future review.

Concluding remarks

The use of AI and RWD has been emerging but focused on limited areas across several stages of the drug development process. Most AI studies focused on AE detection from clinical narratives in EHRs and a few studies explored applications for trial recruitment optimization and clinical drug repurposing. Benefitting from the detailed, longitudinal, multidimensional large collections of RWD and powerful AI algorithms, the use of AI methods on RWD provides golden opportunities in drug development, especially in identifying previously unknown associations and generating new hypotheses. Nevertheless, several current research gaps and challenges exist, such as issues in data quality, the difficult of sharing clinical data, and the lack of interpretability and transportability in AI models. We have highlighted examples of latest advancements in AI and data science that could address these challenges. For example, the increasing capability of DL models that can handle longitudinal and heterogeneous RWD and the raise of causal AI provide new research opportunities in drug development that can benefit from the combined use of AI and RWD.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgment

This work is supported in part by NIH grants UL1TR001427, R01CA246418, R21AG068717, R21CA245858, and R21ES032762; CDC/NIDDK grant U18DP006512; PCORI grants ME-2018C3-14754 and CDRN-1501-26692; and the Cancer Informatics Shared Resource of the University of Florida Health Cancer Center. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH, CDC or PCORI.

References

1. Decker S and Sausville EA et al. (2007) Drug discovery. In Principles of Clinical Pharmacology (2nd edn) (Atkinson AJ, ed.), pp. 439–447, Academic Press
2. McLean L et al. (2015) Drug development. In Rheumatology (6th edn)(Hochberg MC, ed.), pp. 395–400, Elsevier
3. Paul SM et al. (2010) How to improve R&D productivity: the pharmaceutical industry's grand challenge. Nat. Rev. Drug Discovery 9 (3), 203–214 [PubMed: 20168317]
4. Moore TJ et al. (2018) Estimated costs of pivotal trials for novel therapeutic agents approved by the US Food and Drug Administration, 2015–2016. JAMA Int. Med. 178 (11), 1451–1457
5. FDA (2019) Use of Real-World Evidence to Support Regulatory Decision-Making for Medical Devices. FDA
6. FDA (2019) Promoting Effective Drug Development Programs: Opportunities and Priorities for FDA's Office of New Drugs. FDA
7. FDA (2020) Real-World Evidence. FDA
8. Sherman RE et al. (2016) Real-world evidence—what is it and what can it tell us? N. Engl. J. Med. 375 (23), 2293–2297 [PubMed: 27959688]
9. FDA (2020) Submitting Documents Using Real-World Data and Real-World Evidence to FDA for Drugs and Biologics Guidance for Industry. FDA

10. Vamathevan J et al. (2019) Applications of machine learning in drug discovery and development. Nat. Rev. Drug Discovery 18 (6), 463–477 [PubMed: 30976107]
11. Jeon J et al. (2014) A systematic approach to identify novel cancer drug targets using machine learning, inhibitor design and high-throughput screening. Genome Med. 6 (7), 57 [PubMed: 25165489]
12. Ferrero E et al. (2017) In silico prediction of novel therapeutic targets using gene-disease association data. J. Transl. Med. 15 (1), 182 [PubMed: 28851378]
13. Vafaee F et al. (2018) A data-driven, knowledge-based approach to biomarker discovery: application to circulating microRNA markers of colorectal cancer prognosis. NPJ Syst. Biol. Appl. 4, 20 [PubMed: 29872543]
14. FDA (2020) The Drug Development Process. FDA
15. Jiang F et al. (2017) Artificial intelligence in healthcare: past, present and future. Stroke Vasc. Neurol. 10.1136/svn-2017-000101 Published online 12 20, 2017
16. Davenport T and Kalakota R (2019) The potential for artificial intelligence in healthcare. Fut. Healthcare J. 6 (2), 94–98
17. Singh G et al. (2018) Real world big data for clinical research and drug development. Drug Discovery Today 23 (3), 652–660 [PubMed: 29294362]
18. Tricco A et al. (2020) A scoping review of rapid review methods. BMC Med. 13, 224
19. Dobbins M (2017) Rapid Review Guidebook. National Collaborating Centre for Methods and Tools
20. Aronson JK (2004) What is a clinical trial? Br. J. Clin. Pharmacol. 58, 1–3
21. Lai YS and Afseth JD (2019) A review of the impact of utilising electronic medical records for clinical research recruitment. Clin. Trials 16 (2), 194–203 [PubMed: 30764659]
22. Warrer P et al. (2012) Using text-mining techniques in electronic patient records to identify ADRs from medicine use. Br. J. Clin. Pharmacol. 73 (5), 674–684 [PubMed: 22122057]
23. Luo Y et al. (2017) Natural language processing for EHR-based pharmacovigilance: a structured review. Drug Safety 40 (11), 1075–1089 [PubMed: 28643174]
24. Christopoulou F et al. (2020) Adverse drug events and medication relation extraction in electronic health records with ensemble deep learning methods. JAMIA 27 (1), 39–46 [PubMed: 31390003]
25. Zhao J et al. (2015) Predictive modeling of structured electronic health records for adverse drug event detection. BMC Med. Inf. Decis. Making 15 (Suppl. 4), S1
26. Pfaff E et al. (2019) Recruiting for a pragmatic trial using the electronic health record and patient portal: successes and lessons learned. J. Am. Med. Inform. Assoc. 26 (1), 44–49 [PubMed: 30445631]
27. Embi PJ et al. (2005) Development of an electronic health record-based clinical trial alert system to enhance recruitment at the point of care. AMIA Annu. Symp. Proc. 2005, 231–235
28. Tissot HC et al. (2020) Natural language processing for mimicking clinical trial recruitment in critical care: a semi-automated simulation based on the LeoPARDS trial. IEEE J. Biomed. Health Inf. 24, 2950–2959
29. Yang X et al. (2020) Clinical concept extraction using transformers. J. Am. Med. Inform. Assoc. 27, 1935–1942 [PubMed: 33120431]
30. Fu S et al. (2020) Clinical concept extraction: a methodology review. J. Biomed. Inf. 109, 103526
31. Yang X et al. (2019) MADEx: a system for detecting medications, adverse drug events, and their relations from clinical notes. Drug Safety 42 (1), 123–133 [PubMed: 30600484]
32. Zhang H et al. (2017) Continued statin prescriptions after adverse reactions and patient outcomes: a cohort study. Ann. Internal Med. 167 (4), 221–227 [PubMed: 28738423]
33. Spasic I et al. (2019) Cohort selection for clinical trials from longitudinal patient records: text mining approach. JMIR Med. Inf. 7 (4), e15980
34. Xu H et al. (2015) Validating drug repurposing signals using electronic health records: a case study of metformin associated with reduced cancer mortality. J. Am. Med. Inf. Assoc. 22 (1), 179–191
35. Kuang Z et al. (2019) A machine-learning-based drug repurposing approach using baseline regularization. Methods Mol. Biol. 1903, 255–267 [PubMed: 30547447]

36. Xu H et al. (2020) Electronic health records for drug repurposing: current status, challenges, and future directions. Clin. Pharmacol. Ther. 107 (4), 712–714 [PubMed: 32012237]
37. Meystre SM et al. (2008) Extracting information from textual documents in the electronic health record: a review of recent research. Yearbook Med. Inf. 2008, 128–144
38. Boland MR et al. (2013) Defining a comprehensive verotype using electronic health records for personalized medicine. J. Am. Medi. Inf. Assoc. 20 (e2), e232–e238
39. Pearl J and Bareinboim E (2011) Transportability of causal and statistical relations: a formal approach. 2011 IEEE 11th International Conference on Data Mining Workshops 540–547 IEEE
40. Bareinboim E and Pearl J (2016) Causal inference and the data-fusion problem. Proc. Natl. Acad. Sci. U. S. A. 113 (27), 7345–7352 [PubMed: 27382148]
41. Li H and Fan Y (2019) Early prediction of Alzheimer's disease dementia based on baseline hippocampal MRI and 1-year follow-up cognitive measures using deep recurrent neural networks. Proc. IEEE Int. Symp. Biom. Imaging 2019, 368–371
42. Baytas IM et al. (2017) Patient subtyping via time-aware LSTM networks. In Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining 65–74 Association for Computing Machinery
43. Li Y et al. (2020) BEHRT: transformer for electronic health records. Sci. Rep. 10 (1), 7155 [PubMed: 32346050]
44. Wu Y et al. (2018) Combine factual medical knowledge and distributed word representation to improve clinical named entity recognition. AMIA Annu. Symp. Proc. 2018, 1110–1117 [PubMed: 30815153]
45. Pearl J (2019) The seven tools of causal inference, with reflections on machine learning. Commun. ACM 62 (3), 54–60
46. Pittavino M et al. (2017) Comparison between generalized linear modelling and additive Bayesian network; identification of factors associated with the incidence of antibodies against *Leptospira interrogans* sv *pomona* in meat workers in New Zealand. Acta Tropica 173, 191–199 [PubMed: 28487178]
47. Andra SS et al. (2015) Preliminary evidence of the association between monochlorinated bisphenol A exposure and type II diabetes mellitus: a pilot study. J. Environ. Sci. Health. Part A, Toxic/ Hazard. Subst. Environ. Eng. 50 (3), 243–259
48. Lewis FI and McCormick BJJ (2012) Revealing the complexity of health determinants in resource-poor settings. Am. J. Epidemiol. 176 (11), 1051–1059 [PubMed: 23139247]
49. Hernán MA and Robins JM (2016) Using Big Data to emulate a target trial when a randomized trial is not available. Am. J. Epidemiol. 183, 758–764 [PubMed: 26994063]
50. Payrovnaziri SN et al. (2020) Explainable artificial intelligence models using real-world electronic health record data: a systematic scoping review. J. Am. Med. Inf. Assoc. 27, 1173–1185
51. Collins FS et al. (2014) PCORnet: turning a dream into reality. J. Am. Med. Inf. Associ. 21 (4), 576–577
52. Hripcsak G et al. (2015) Observational Health Data Sciences and Informatics (OHDSI): opportunities for observational researchers. Stud. Health Technol. Inf. 216, 574–578
53. Visweswaran S et al. (2018) Accrual to Clinical Trials (ACT): A Clinical and Translational Science Award Consortium Network. JAMIA Open 1 (2), 147–152 [PubMed: 30474072]
54. Yang X et al. (2019) A study of deep learning methods for de-identification of clinical notes in cross-institute settings. BMC Med. Inf. Decis. Making 19 (5), 232
55. Bian J et al. (2019) Implementing a hash-based privacy-preserving record linkage tool in the OneFlorida clinical research network. JAMIA Open 2 (4), 562–569 [PubMed: 32025654]
56. Holford N et al. (2010) Clinical trial simulation: a review. Clin. Pharmacol. Ther. 88 (2), 166–182 [PubMed: 20613720]
57. Bowton E et al. (2014) Biobanks and electronic medical records: enabling cost-effective research. Sci. Transl. Med. 6 (234), 234cm3
58. Gns HS et al. (2019) An update on drug repurposing: re-written saga of the drug's fate. Biomed. Pharmacother. 110, 700–716 [PubMed: 30553197]

59. Pappa D and Stergioulas LK (2019) Harnessing social media data for pharmacovigilance: a review of current state of the art, challenges and future directions. Int. J. Data Sci. Anal. 8 (2), 113–135

60. Tricco AC et al. (2018) Utility of social media and crowd-intelligence data for pharmacovigilance: a scoping review. BMC Med. Inf. Decis. Making 18 (1), 38

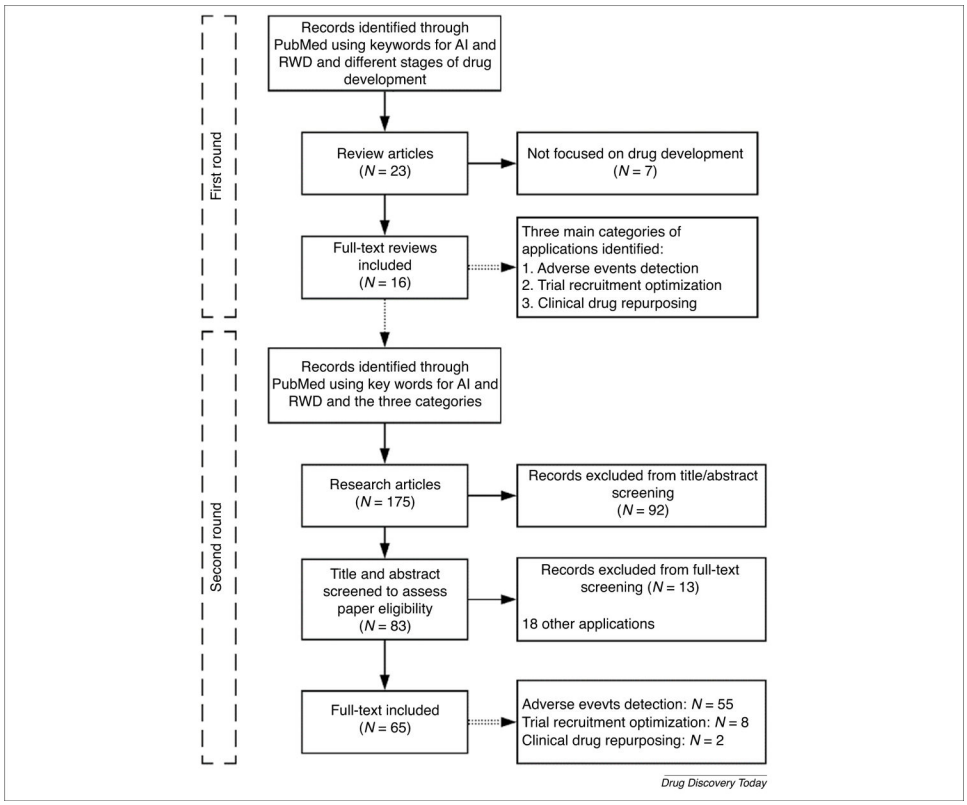

FIGURE 1.
The overall search and screening proces

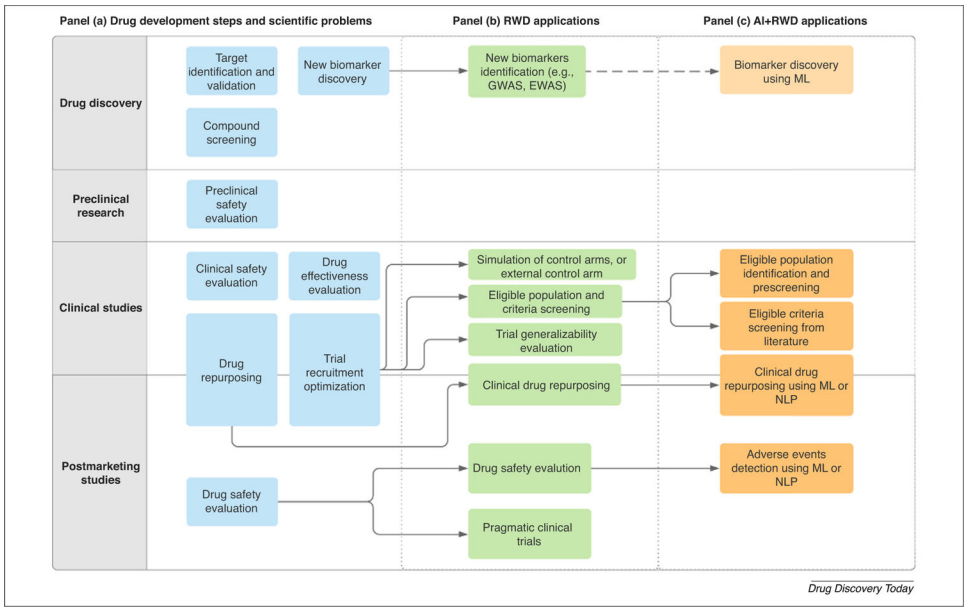

FIGURE 2.

Identified artificial intelligence (AI) and real-world data (RWD) applications across the different stages in the drug development process. Abbreviations: EWAS, epigenome-wide association study; GWAS, genome-wide association study; ML, machine learning.

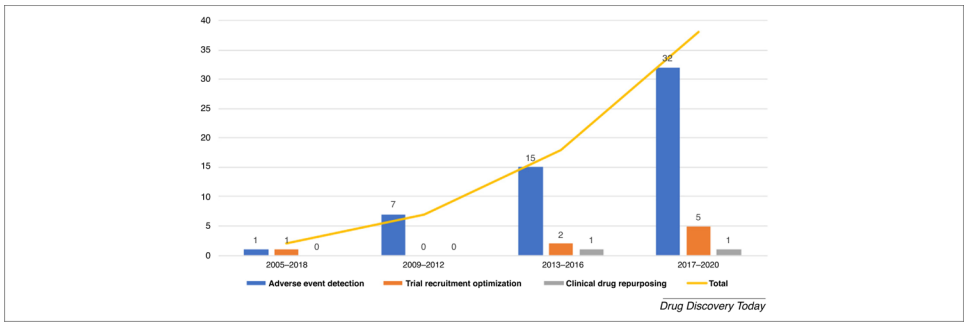

FIGURE 3.

Number of original studies with artificial intelligence (AI) methods using real-world data (RWD) in the drug development process over the years.

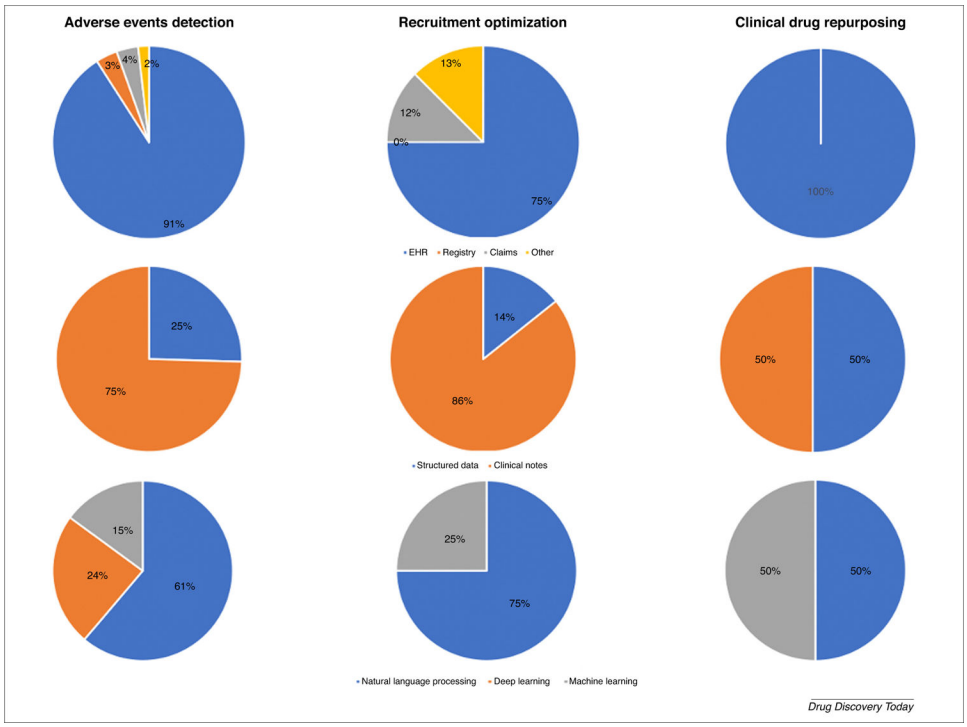

FIGURE 4.

Breakdown of real-world data sources, data types, and artificial intelligence (AI) methods used in the identified applications across the drug development process. Given the overwhelming number of studies used AI-driven natural language processing (NLP) methods, we separated NLP studies from other machine/ deep-learning (ML/DL) studies.

TABLE 1

The main categories of AI and RWD applications in drug development

Applications	Subcategories	Examples	Refs
AE (adverse event) detection	Mining clinical notes using NLP (natural language processing)	DL-based NLP to detect AEs in clinical notes extracted from EHR	[24]
	Mining structured EHR (elctronic health record) data	Predictive modeling of structured EHRs for AE detection	[25]
Recruitment optimization	Electronic recruitment through EHR	Electronic recruitment integrated into EHR workflow that sends electronic messages to recruit eligible patients	[26,27]
	Eligible population identification/prescreening	Automated review of EHR to identify eligible population using NLP	[28]
Clinical drug repurposing		Comparison between diabetic and nondiabetic patients with cancer showed that use of metformin was associated with decreased mortality after cancer diagnosis	[34]

Drug Discov Today. Author manuscript; available in PMC 2021 November 27.