Real world big data for clinical research and drug development

Gurparkash Singh[1], Duane Schulthess[2], Nigel Hughes[3], Bart Vannieuwenhuyse[3] and Dipak Kalra[4]

[1] Janssen Research and Development, Fremont, CA, USA
[2] Vital Transformation, 107 Leopold III Laan, Wezembeek-Oppem 1970, Belgium
[3] Janssen Research and Development, Beerse, Belgium
[4] Dept. Medical Informatics & Statistics, University of Ghent, De Pintelaan 185, Gent 9000, Belgium

The objective of this paper is to identify the extent to which real world data (RWD) is being utilized, or could be utilized, at scale in drug development. Through screening peer-reviewed literature, we have cited specific examples where RWD can be used for biomarker discovery or validation, gaining a new understanding of a disease or disease associations, discovering new markers for patient stratification and targeted therapies, new markers for identifying persons with a disease, and pharmacovigilance. None of the papers meeting our criteria was specifically geared toward new novel targets or indications in the biopharmaceutical sector; the majority were focused on the area of public health, often sponsored by universities, insurance providers or in combination with public health bodies such as national insurers. The field is still in an early phase of practical application, and is being harnessed broadly where it serves the most direct need in public health applications in early, rare and novel disease incidents. However, these exemplars provide a valuable contribution to insights on the use of RWD to create novel, faster and less invasive approaches to advance disease understanding and biomarker discovery. We believe that pharma needs to invest in making better use of EHRs and the need for more precompetitive collaboration to grow the scale of this 'big denominator' capability, especially given the needs of precision medicine research.

Introduction

Access to large-scale real-world data (RWD) to support basic and translational science in clinical research and development is a significant opportunity and challenge for life sciences and the pharmaceutical industry. It is well recognized that randomized, controlled trials provide high-quality data on restricted patient populations (little co-morbidity, not including older patients, etc.) and that high-volume, routinely collected data have the potential to provide insights into the health situation and treatment effectiveness in a more representative diversity of patients, as well as to permit hypothesis generation regarding rare conditions, rare effects and rare biomarkers. Population health data have been used as a source of knowledge discovery for decades, therefore using RWD in analysis is not new. Numerous population cohorts, usually operating on a national or regional basis, and usually with many thousands of patients each, have generated a vast body of epidemiological literature. Disease, procedure and other health registries, often curated at national or regional levels, have similarly resulted in an expansive volume of scientific literature. At the opposite end of the RWD spectrum, individual care organizations such as hospitals and general practitioners (GPs) have long used their locally held data for quality and safety monitoring (e.g., via audit), and many have established clinical data warehouses for internal research use. The incorporation of electronic health

Corresponding author: Kalra, D. (dipak.kalra@eurorec.org)

https://doi.org/10.1016/j.drudis.2017.12.002

record (EHR) data for research at a care site level is now well recognized and supported [1,2]. Claims databases are widely available on a large scale and are used for population health research, but have recognized selection and up-coding bias that questions the scientific validity of real world evidence (RWE) derived from them [3,4]. The focus of this review is on the novel large-scale use of routinely collected health record data; therefore, claims databases were not included.

Aggregations of data across multiple organizations also exist in practice and this has proven to be a valuable scientific approach to working with RWD. One of the best known in Europe is the Clinical Practice Research Datalink (CPRD) [5], a governmental, not-for-profit, research service, jointly funded by the NHS National Institute for Health Research (NIHR) and the Medicines and Healthcare products Regulatory Agency (MHRA) in the UK. Providing anonymized primary care records for public health research since 1987, research using CPRD data has resulted in >1700 publications in drug safety, best practice and clinical guidelines [6]. As another example, the Italian Medicines Agency, Agenxia Italiana del Farmaco (AIFA), has set up a system of registries for RWD collection as part of the reimbursement and pricing process to ensure the licensed medicines meet pre-agreed effectiveness targets [7]. According to this publication, there are currently >120 registries, through which >80 medicines are monitored in >50 therapeutic indications. Different stakeholders have different access rights to the system across 21 regions, >1000 hospitals, >24 000 clinicians, 1500 pharmacists and 32 marketing authorization holders.

A further important contribution to the conduct of research on big health data is the Observational Health Data Sciences and Informatics (OHDSI) collaboration, which enables the scaling up of research through the adoption of common data model and tools. For example, Hripcsak *et al.* used OHDSI to combine data from 11 data sources, a total of 250 million patients, to examine treatment pathways in type 2 diabetes mellitus, hypertension and depression [8]. Substantial R&D investments are currently being made to develop tools, platforms and governance processes to enable the distributed analysis of multiple EHR systems [9,10]. One of the most ambitious projects is the 5-year, €56 million EU-funded European Medical Information Framework (EMIF), which is a multi-stakeholder platform creating an EU technology and governance framework that will enable the re-use and management of existing health data [11].

The objective of this paper is to identify the extent to which RWD is being utilized, or could be utilized, at scale in drug discovery, such as the identification and targeting of novel therapeutic areas. Through a comprehensive screening of peer-reviewed literature, we have cited specific examples where RWD can be used for biomarker discovery or validation, gaining a new understanding of a disease or disease associations, discovering new markers for patient stratification and targeted therapies, new markers for identifying persons with a disease and pharmacovigilance. In the context of this article, the term 'real world data' is used to describe data sources that are collected or measured outside of the randomized, controlled trial, and reflective of clinical management or naturalistic care [12]. These can include cohort studies, patient registries and data generated by patients directly [13,14]. The growing body of data held within high quality EHR systems, and the adoption of interoperability standards and harmonization methodologies, has made the large-scale analysis of EHR data more attractive and viable to aid in the development of needed new therapies [15]. To our knowledge, there has been no formal literature review examining the use of RWD sources to successfully generate new evidence in support of drug development. There is no consensus definition of big health data and, in the context of this article, we have opted to define 'big' as analyzing the data on 1 million or more subjects within RWD sets, either in one dataset or distributed over several datasets, to profile a relevant subpopulation for the published research (we later discuss the limitations of this definition).

We have sought empirical studies that have required a large population denominator to identify sufficient relevant patient numbers to generate robust results. We recognize that this inclusion criterion is not based on any authoritative or widely adopted definition or convention, and we hope that, by examining the success of evidence generation based on this definition, we will stimulate community debate on how the use of RWD for clinical research and drug development should best be characterized and differentiated from well-established epidemiological and health services research uses of data sources such as registries and CPRD. We reviewed the literature using the search string with different inclusion terms and a keyword: 'million', which was hard-wired into the title or abstract to capture the scale of study that we would consider to be suitably 'big' (Fig. 1). It was a deliberate decision to search for specific mention of a large dataset size, because our preliminary exploration of the literature revealed many studies that were conventional in scale but utilized terms like 'big data' rather indiscriminately. However, we recognize that the RWD community still needs to agree on a precise term that could be used for future literature reviews of this kind, as discussed later, especially when considering rare conditions where a large population database is required to find relatively small numbers of precisely specified patients. Without setting any date limitations, the above search string identified 534 publications in PubMed. However, as can be seen in Fig. 2, the clear majority of the publications that were screened were published within the past 10 years. We adopted a manual title and abstract screening to characterize the retrieved results against those kinds of evidence that are most relevant to clinical research and development, and the subject of this literature review. A total of 32 publications were retained, and subjected to independent full-paper review by three authors. Given our focus on RWE supporting clinical research and development, our full-paper screening sought to verify our initial inclusion criteria and select only those papers with novel techniques and findings that were directly applicable to current needs of biopharmaceutical R&D on the basis of five kinds of evidence, under which our findings are grouped. Twenty publications were retained for inclusion in this review; see Fig. 3 for the PRISMA diagram [16]. Most of the publications we eliminated in screening were describing the potential of RWD as an opportunity, sometimes with examples of knowledge gaps that might be filled through RWD. However, this large body of publications did not offer any actual findings from data. Our screening clearly demonstrated a diminishing number of studies reporting the practical application of RWD to drug development as one went back in time. The greatest

Please cite this article in press as: Singh, G. et al. Real world big data for clinical research and drug development, Drug Discov Today (2018), https://doi.org/10.1016/j.drudis.2017.12.002

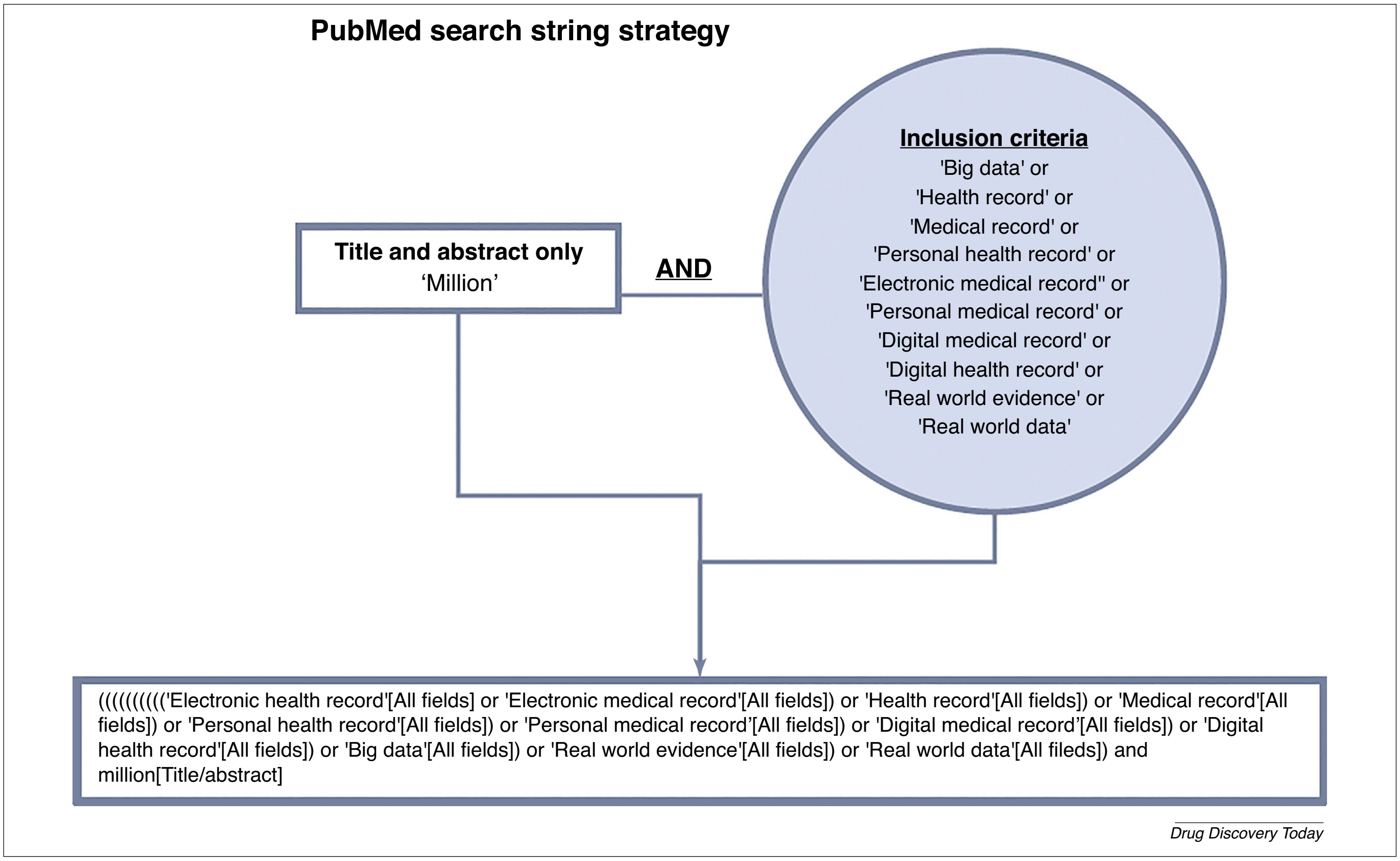

PubMed search string strategy
Title and abstract only
'Million'
AND
Inclusion criteria
'Big data' or
'Health record' or
'Medical record' or
'Personal health record' or
'Electronic medical record'' or
'Personal medical record' or
'Digital medical record' or
'Digital health record' or
'Real world evidence' or
'Real world data'
(((((((((('Electronic health record'[All fields] or 'Electronic medical record'[All fields]) or 'Health record'[All fields]) or 'Medical record'[All fields]) or 'Personal health record'[All fields]) or 'Personal medical record'[All fields]) or 'Digital medical record'[All fields]) or 'Digital health record'[All fields]) or 'Big data'[All fields]) or 'Real world evidence'[All fields]) or 'Real world data'[All fileds]) and million[Title/abstract]
Drug Discovery Today

FIGURE 1

Graphic representation of search strategy. Literature was reviewed by combining a set of inclusion criteria for kinds of health record source, combined only with the keyword 'million' to generate the actual search string. The term million was hard-wired into the title or abstract as an indication of the scale of study that we would consider to be suitably 'big'.

alignment to the objectives of the study was demonstrated in the past five years, where we limited our screening for final inclusion.

During the manual title and abstract screening, we excluded publications that described the needs, opportunities or challenges of using big data, or that described databases that had the potential to be used for big data research but did not include any concrete empirical research findings. We also excluded editorials and publications describing big data research methodologies without offering any empirical findings. Studies that were literature reviews, not empirical studies, were used for source material and relevant articles that fitted our criteria were incorporated into the abstract search and screening, but the review articles themselves were not included unless they reported original empirical findings. The 20 selected studies fall into five basic applications of RWD for clinical research and development (Fig. 4). Consensus was reached on any papers that did not have a unanimous decision at a dedicated face-to-face meeting. Details about the size and source of the datasets are presented in Table 1. Below, we have included papers that provide a concrete example of a use of RWE that could be harnessed and repurposed for drug discovery. We give an in-depth background on one example that we feel best exemplifies a use case for RWE, and also highlight several other examples that met our criteria.

Case study

Biomarker discovery or validation

Only one publication highlighted novel uses in the development or application of RWD techniques in the identification or validation of new biomarkers. Of particular interest was Zodiac, the use of a Bayesian model to create a more effective map of cancer outcomes based on the analysis of genetic interactions as biomarkers found in the TCGA database of 200 million patient records [17].

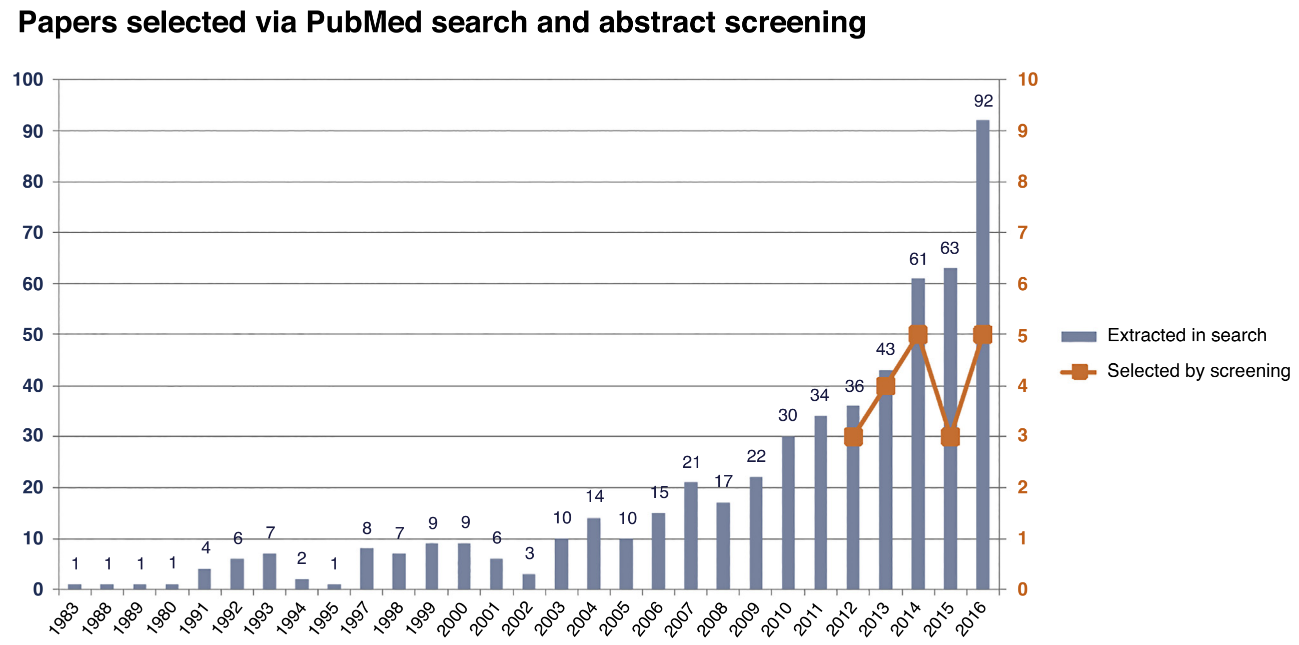
A new understanding of a disease or a disease association

Six of our selected papers demonstrate how novel uses of RWD can foster new understandings of disease associations and or comorbidities that would be particularly useful when trying to target new populations or indications for research. Of note was a study that used the Taiwan National Health Insurance Database of >782 million outpatient visits to develop the Cancer Associations Map Animation (CAMA). By tracking previously unmapped cancer–disease associations across ages and genders, CAMA can effectively detect cancer comorbidities earlier than is possible by manual inspection and identify potential effect modifiers or new risk factors [18].

Other studies:

- An analysis of 25 million patient records of the US Veterans Administration discovered that those with periodontal disease were more likely to have rheumatoid arthritis [19].

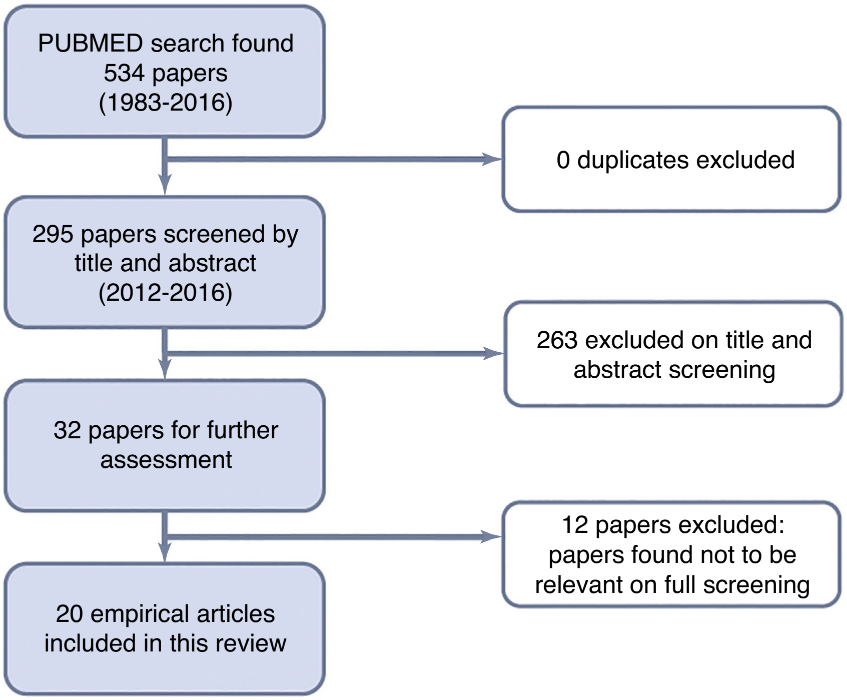
Please cite this article in press as: Singh, G. et al. Real world big data for clinical research and drug development, Drug Discov Today (2018), https://doi.org/10.1016/j.drudis.2017.12.002



Drug Discovery Today

FIGURE 2

Number of publications by year (since 1983) retrieved using the search strategy (left hand axis), and number retained after title and abstract screening (right hand axis). Three of the authors (G.S., D.S. and D.K.) contributed to independent title and abstract screening of publications in reverse chronological order (i.e., starting with 2016), discussing via regular teleconferences any publications that did not have a three-way consensus on the decision. Note: online versions of two papers ([32] and [33]) were available in 2016.


PUBMED search found 534 papers (1983-2016)
0 duplicates excluded
295 papers screened by title and abstract (2012-2016)
263 excluded on title and abstract screening
32 papers for further assessment
12 papers excluded: papers found not to be relevant on full screening
20 empirical articles included in this review

Drug Discovery Today

FIGURE 3

Preferred reporting items for systematic reviews and meta-analyses (PRISMA) diagram for this review.

Please cite this article in press as: Singh, G. et al. Real world big data for clinical research and drug development, Drug Discov Today (2018), https://doi.org/10.1016/j.drudis.2017.12.002

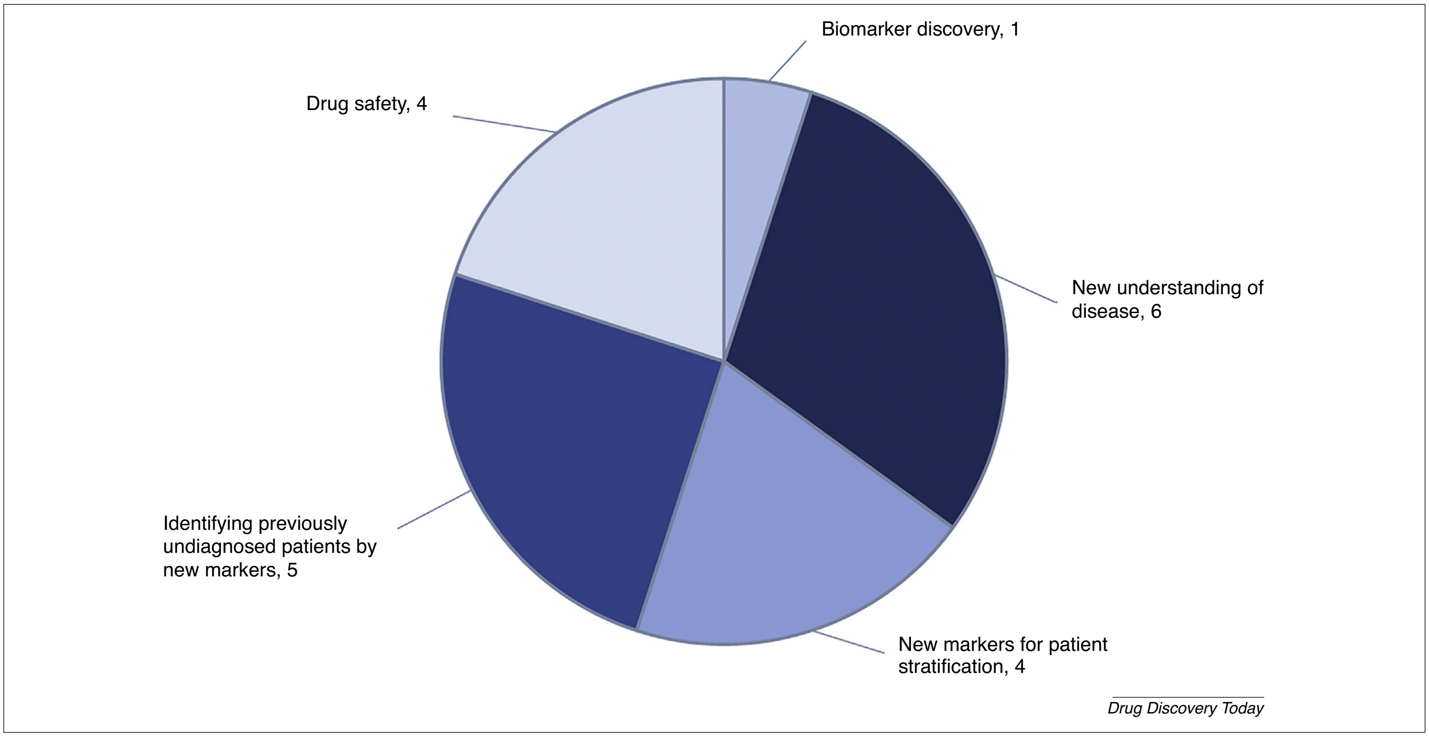


FIGURE 4

Distribution of selected studies among the five categories.

The long-held belief that an increase in the risk of mortality is encountered using long-acting β_2-agonist (LABA) monotherapy in the treatment of asthma could not be proven in a RWD study of a cohort of 994 627 patients [20].

- Primary open angle glaucoma (POAG) is associated to a moderate increase in risk for vascular dementia. Further, the likelihood of a hospital record of POAG following Alzheimer's disease or vascular dementia was very low [21].
- Linking EHRs in CALIBER (cardiovascular research using linked bespoke studies and electronic health records) found the assumption that blood pressure has an impact on all cardiovascular diseases, and diastolic and systolic associations are concordant, which is not supported by outcomes data [22].
- By modeling temporal relationships between 41.2 million time-stamped international classifications of diseases in 1.6 million patients, researchers discovered that diabetes usually preceded the diagnosis of *Helicobacter pylori* (bacteria linked ulcers), leading to questions of cause and effect of the two conditions [23].

Discovering or validating new markers for patient stratification and targeted therapies

With the continuing push toward stratified, targeted therapies, the use of RWD has immediate implications for drug development and efficacy, and our research identified four examples where the use of large datasets created novel approaches to stratification of patient populations. One of the selected studies developed a novel approach that could be generalized across multiple disease areas. By using a flexible framework called generalized low rank models (GLRM), the researchers could successfully capture known and putative phenotypes using vastly different datasets including text from physician notes [24].

Other studies selected included:

- The use of RWD to inform and verify the use of concomitant corticosteroid in the treatment of patients with metastatic castration-resistant prostate cancer [25].
- A study of 27 million patient records that accurately determined individual risk factors post knee arthroplasty [26].
- An analysis of EHR data taken from multiple healthcare systems over the period 1999–2011 found that patient weight had more effect than height on venous thromboembolic events [27].

New markers for identifying persons with a disease (e.g., formerly undiagnosed patients)

Drug development will increasingly require identification of new disease markers that can better identify previously undiagnosed patients, and our research found five examples of this. Specifically, given the lack of treatments in neurological disorders, the use of algorithms in the identification of new patents is a pressing need for the biopharmaceutical sector. One study outlined the effective application of semiautomated mining of EHRs to ascertain bipolar disorder patients and control subjects with high specificity and predictive value when compared with diagnostic interviews [28]. This technique could have broad applicability across many research areas in neurology.

Other studies selected included:

- The outcomes of 2.8 million data points taken from the real world pragmatic use of the therapy ranibizumab in the treatment of age-related macular degeneration when compared with the results of the randomized clinical trial [29].

Please cite this article in press as: Singh, G. et al. Real world big data for clinical research and drug development, Drug Discov Today (2018), https://doi.org/10.1016/j.drudis.2017.12.002

TABLE 1

Overview of selected articles

Publication year	Article title	Data size	Data sources	RWD for clinical R&D	Refs
2015	Zodiac: a comprehensive depiction of genetic interactions in cancer by integrating TCGA data	200 million pairs of genes	The Cancer Genome Atlas	Biomarker discovery	[17]
2016	Cancer-disease associations: a visualization and animation through medical big data	782 million	Taiwan National Health Insurance Database	New understanding of disease	[18]
2015	Using big data to evaluate the association between periodontal disease DNA rheumatoid arthritis	25 million patients	US Veterans Health Administration Repository (1999–2012)	New understanding of disease	[19]
2016	Pharmacoepidemiological study of long-acting β-agonist/inhaled corticosteroid therapy and asthma mortality	About 1 million asthma patients	Asthma Safety Observational Study – EH records from ten collaborating institutions (Jan 2000–Dec 2010)	New understanding of disease	[20]
2015	Associations between primary open angle glaucoma, Alzheimer's disease and vascular dementia: record linkage study	2.5 million reference cohort	English National Health Service linked hospital episode statistics from 1999 to 2011	New understanding of disease	[21]
2014	Blood pressure and incidence of 12 cardiovascular diseases: lifetime risks, healthy life-years lost, and age-specific associations in 1.25 million people	1·25 million patients	CALIBER (Cardiovascular research using linked bespoke studies and electronic health records)	New understanding of disease	[22]
2014	Modeling temporal relationships in large-scale clinical associations	41.2 million time-stamped International Classification of Diseases, Ninth Revision (ICD-9) codes	1.6 million patients	New understanding of disease	[23]
2016	Discovering patient phenotypes using generalized low rank models	8 million hospitalization records	2010 Healthcare Cost and Utilization Project National Inpatient Sample	Stratification	[24]
2013	Real-world corticosteroid utilization patterns in patients with metastatic castration-resistant prostate cancer in 2 large US administrative claims databases	31 million individuals	Two large USA databases	Stratification	[25]
2014	Risk factors for manipulation after total knee arthroplasty: a pooled electronic health record database study	27 million patients	Explorys (Explorys, Inc., Cleveland, OH) was used to mine a pooled electronic healthcare database	Stratification	[26]
2012	Patient characteristics associated with venous thromboembolic events: a cohort study using pooled electronic health record data	0.95 million patients	1999 to 2011 EHR data from multiple healthcare systems	Stratification	[27]
2015	Validation of electronic health records phenotyping of bipolar disorder and controls	4.2 million patients	Partners Healthcare Research Patient Data Registry	Disease identification	[28]
2014	The neovascular age-related macular degeneration database: multicenter study of 92 976 ranibizumab injections: report 1: visual acuity	2.8 million data points from 300 000 clinical visits	14 UK centers to a central database using an electronic medical record (EMR) system	Disease identification	[29]
2013	Pulmonary embolism, myocardial infarction, and ischemic stroke in lung cancer patients: results from a longitudinal study	3 million hospitalizations	Dutch PHARMO medical record linkage system	Disease identification	[30]
2013	Incidence and mortality of acute and chronic pancreatitis in The Netherlands: a nationwide record-linked cohort study for the years 1995–2005	18 million Dutch citizens	Dutch hospital databases	Disease identification	[31]
2017	Developing electronic health record algorithms to accurately identify patients with systemic lupus erythematosus	2.5 million subjects	Vanderbilt Synthetic Derivative	Disease identification	[32]
2017	Exploring trends of nonmedical use of prescription drugs and polydrug abuse in the Twittersphere using unsupervised machine learning	11 million tweets	Microblogging site Twitter	Drug safety	[33]
2014	Using aggregated, de-identified electronic health record data for multivariate pharmacosurveillance: a case study of azathioprine	10 million individuals	Explore platform (Explorys, Inc. now IBM Watson)	Drug safety	[34]
2012	Surveillance for Guillain–Barré syndrome after influenza vaccination among the Medicare population, 2009–2010	14.0 million vaccination records	Medicare Claims Data	Drug safety	[35]
2012	Using temporal patterns in medical records to discern adverse drug events from indications	9 million clinical notes for more than 1 million patients	Stanford Clinical Data Warehouse (STRIDE)	Drug safety	[36]

Please cite this article in press as: Singh, G. et al. Real world big data for clinical research and drug development, Drug Discov Today (2018), https://doi.org/10.1016/j.drudis.2017.12.002

- The risk of lung cancer patients developing pulmonary embolism when compared with cancer-free controls when analyzing 3 million Dutch hospitalizations [30].
- A nationwide cohort study of the incidence and mortality of acute and chronic pancreatitis in The Netherlands found that disease burden and healthcare costs will probably increase, linked to the ageing Dutch population [31].
- An algorithm developed at Vanderbilt University that enabled the rapid searching of an EHR database of 2.5 million subjects to accurately identify systemic lupus erythematosus [32]. The ability to use algorithms and large datasets to rapidly identify previously undiagnosed and unknown patient populations would not only have a direct impact on lupus research but also has the potential to be applicable to autoimmune disorders more broadly.

Drug safety studies

Several examples of RWD can be found in the application of drug safety. Naturally, speed and accuracy of discovery are of vital importance in safety-related situations. An accurate understanding of adverse events would be of enormous benefit to regulators, patients and industry. The ability to utilize social media to automate drug safety monitoring could radically reduce its costs and accelerate results and we selected four papers as exemplars of best practices. The most unique of these explored the practical use of 11 million 'Tweets' to determine the frequency of prescription drug and polydrug abuse using unsupervised machine learning. The study concluded that social media could be a viable methodology for drug abuse surveillance [33].

Other studies selected:

- Aggregated, de-identified EHR data for multivariate pharmacosurveillance of 10 million individuals could provide sufficient insight and statistical power to detect potential patterns of medication side-effect associations [34].
- Claims-based surveillance of >14 million vaccinations did not indicate a statistically significant elevated Guillain–Barré syndrome rate following seasonal or H1N1 influenza vaccination [35].
- Nine million clinical notes for >1 million patients were used to detect statistically significant drug safety signals at co-occurrences of drug–disease mentions [36].

Discussion

Whereas our papers selected were all in unique areas of clinical applications, there are several overarching themes that they share. First, the use of large datasets broadly enables a far better understanding of treatment pathways in diagnosis and efficiency of treatment, as well as drug safety. Second, with the ability to harness multiple EHR systems, it is now possible to sift for rare indications, and develop unique algorithms to find therapeutic 'diamonds in the rough', as well as uncover previously missed or early indications of disease incidents that might have previously been undetectable without the judicious harnessing of RWE. The fact is, although there is a tsunami of sky-high rhetoric related to big data being promulgated, our selected papers show that this work is still in an early phase of practical application, and is being harnessed broadly where it serves the most direct need in public health applications in early, rare and novel disease incidents. RWE is delivering results, but it is not yet ubiquitous outside of a few areas in public health. Additionally, one of our key questions this paper set to answer, that RWE can be used to assist in the targeting of novel therapeutic areas in drug development, has yet to be supported in the papers we have selected. None of the papers we finally identified was specifically geared toward new novel targets or indications in the biopharmaceutical sector. The majority of the studies were focused more generally in the area of public health, often sponsored by the universities themselves, insurance providers or in combination with public health bodies such as national insurers. Given that the current ownership of large public health data is often at the hospital system or national level, this does make sense in hindsight. Much of the usable RWD is housed in large EHRs owned by public health bodies or insurance organizations responsible for reimbursement. It stands to reason that the goals of most public EHR owners are not currently focused in the discovery and development of new molecular entities in the pharmaceutical sector, and could be a reason why our initial goal of finding best-case examples for drug targets has gone unmet by this exercise. As well, given our search strategy has been focused on publicly listed peer-reviewed literature, studies that were business-driven and pharma-sponsored could be largely unpublished and treated as commercially confidential intelligence even if the source data are widely accessible.

Because many of the contributors to this paper are currently collaborating with several industry partners in the use of RWE for drug discovery applications, we do know that the research is occurring but it apparently is not yet appearing in the open body of knowledge as peer-reviewed literature. Given this lack of specific drug discovery examples in the final papers of our screening, we chose studies that clearly demonstrated uses of RWD techniques or applications that can be re-appropriated or reverse-engineered for commercial, unmet medical needs in clinical research and drug development, often in areas that drug companies are currently focused; namely, oncology, neurological disorders, cardiovascular disease (CVD) and autoimmune diseases. As highlighted in our results, any discoveries employing large datasets will need to be investigated to minimize confounding variables and establish their clinical validity to pharmaceutical applications. We noted that data quality was rarely discussed but it is an important consideration for RWD. However, we are confident that these exemplars provide a valuable contribution to insights on the use of RWD to advance disease understanding and biomarker discovery.

Strengths and limitations

There are several important limitations to this literature review that might have impacted the findings. Owing to the rapidly evolving nature of technology, we limited our results to papers from 2012 to 2016 to capture the state of the art; this could exclude some relevant earlier examples. Our EHR search is limited to health records, and excludes other databases such as genomic data, immunochemistry and claims databases. As well, our search only sought evidence within the peer-reviewed literature and there could be examples currently being investigated privately within industrial R&D that are considered proprietary. We therefore recognize that RWD work published in journals listed in PubMed is no more than the tip of the iceberg regarding the use of RWD for drug R&D.

Please cite this article in press as: Singh, G. et al. Real world big data for clinical research and drug development, Drug Discov Today (2018), https://doi.org/10.1016/j.drudis.2017.12.002

In specifically seeking out large-scale uses of RWD (i.e., big data), we limited our data sources to those that had the keyword 'million'. Whereas we assume that 'million' will capture large datasets; we are aware that this might not be the case: the Italian Government and OHDSI examples cited in our introduction were not captured; even though they are best-practice examples in the application of RWD. There is a need for the field to agree on what defines big data and RWD to facilitate consistent empirical research in this area going forward. Discussion is also needed on other measures of signals that should be considered when evaluating a RWD-based study: effect size; number needed to treat; sample size; among others. Consequently, we accept that this literature review offers an exemplary insight rather than a comprehensive examination of the present state of empirical research in the field. We have deliberately not claimed this to be a systematic literature review. We recognize that smaller-scale RWD could also be useful [37].

Implications of this work to pharma

Despite this, a methodology exists for focused literature review that can provide insights for clinical research and drug development pathways utilizing RWD. Targeting of real world studies can elucidate possible partners and collaborators with whom pharmaceutical companies could explore the opportunity to work together on gathering real world insights from their external data sources. In identifying a study, only empirical results are known, and a pharmaceutical researcher will need to establish a partnership to be able, at a minimum, to have an opportunity to review original data (within the ordinary and current constraints of such a task). Beyond this, working within a *quid pro quo* relationship, researchers and original real world study authors have an opportunity to support drug development and work beyond the original author's study remits. It is envisaged this might be a premise for such a collaboration to mutual benefit.

By its very definition, RWD is not necessarily accessible by pharma, requiring local provenance and governance to be protected. Because such remote connectivity to RWD, whether through, for example, federated data networks, common data models as intermediaries or indirect analytical outputs, might be the only agreeable contract with pharma for data custodians of studies as described in this manuscript, such an undertaking is a very different relationship with data than pharma is necessarily conversant with, for instance with its own randomized clinical trial data. Longitudinal collaborations must have a mutual relationship based on trust and transparency of intended use paramount to successful research. To reciprocate with regard to transparency, a call to action is for pharma to expand on adding to the body of evidence in this domain via peer-reviewed publications. As the use of RWE within R&D increases in prominence, evidence will be required not only by pharma as to its veracity but also by regulatory authorities and others who are also needing to understand the role of RWE in 21st century drug development.

Concluding remarks

We have observed a steady, almost exponential, increase in the publication of empirical research that is within the scope of our review. From the early 2000 s we have seen a steady growth in papers about the opportunity of using big health data, methodology papers and papers describing various solutions such as data warehouses and analytics platforms. During the past 5 years, and especially over the past 3 years, we have seen a growing number of actual empirical findings from using big health data relevant to clinical research and drug development. We anticipate continued growth in the quantity, sophistication and scale of this research area.

To accelerate the generation of RWE relevant to clinical research and drug development, we believe that pharma needs to invest in making better use of EHRs and their linkage to molecular databases (within the right governance and technology frameworks). We see the need for more precompetitive collaboration to grow the scale of this 'big denominator' capability, especially given the needs of precision medicine research. We also foresee the need for richer academic-industry-government partnerships, which will depend upon the willingness of governments to provide industry with access to anonymized health data and work collaboratively across academic centers, to reach the necessary population scale. Finally, the authors hope that these opportunities to scale up RWE will help to stimulate improvements in the data quality and interoperability of RWD sources across healthcare and academia.

Conflicts of interest

B.V., N.H. and G.S. are employees of Janssen Research & Development, LLC, and stockholders of Johnson & Johnson.

Acknowledgments

This research has been partly undertaken through the European Medical Informatics Framework project (Grant Agreement No. 115372) funded by the Innovative Medicines Initiative. G.S., a Janssen Bridges Associate, acknowledges the support from his home and host departments, and the Bridges Project Management Office during his Bridges Training. Authors thank Nuala Moran, Anna Ludlow and Petra Naster for their support.

References

1 De Moor, G. *et al.* (2014) Using electronic health records for clinical research: the case of the EHR4CR project. *J. Biomed. Inf.* 53, 162–173

2 Delaney, B.C. *et al.* (2015) Translational medicine and patient safety in Europe: TRANSFoRm–Architecture for the learning health system in Europe. *BioMed. Res. Int.* http://dx.doi.org/10.1155/2015/961526

3 Jensen, E.T. *et al.* (2015) Enrollment factors and bias of disease prevalence estimates in administrative claims data. *Ann. Epidemiol.* 25, 519–525

4 Hyman, J. (2016) The limitations of using insurance data for research. *JADA* 146, 283–285

5 Herrett, E. *et al.* (2015) Data resource profile: Clinical Practice Research Datalink (CPRD). *Int. J. Epidemiol.* 44, 827–836

6 CPRD. Available at: https://www.cprd.com/intro.asp.

7 Montilla, S. *et al.* (2015) Monitoring registries at the Italian Medicines' Agency: fostering access, guaranteeing sustainability. *Int. J. Technol. Assess. Health Care* 31, 210–213

8 Hripcsak, G. *et al.* (2016) Characterizing treatment pathways at scale using the OHDSI network. *PNAS* 113, 7329–7336

9 The Innovative Medicines Initiative 2016. Available at: http://www.imi.europa.eu.

Please cite this article in press as: Singh, G. et al. Real world big data for clinical research and drug development, Drug Discov Today (2018), https://doi.org/10.1016/j.drudis.2017.12.002

10 Kalra, D. *et al.* (2016) The European institute for innovation through health data. *Learn. Health Syst.* 1, 1–8http://dx.doi.org/10.1002/lrh2.10008
11 The European Medical Informatics framework. Available at: http://www.emif.eu.
12 Miani, C. *et al.* Health and Healthcare: Assessing the Real-World Data Policy Landscape in Europe. Available at: https://www.rand.org/randeurope/research/projects/real-world-data-policy-landscape.html.
13 de Lusignan, S. *et al.* (2015) Creating and using real-world evidence to answer questions about clinical effectiveness. *J. Innov. Health Inf.* 22, 368–373
14 Garrison, L.P., Jr *et al.* (2007) Using real-world data for coverage and payment decisions: the ISPOR Real-World Data Task Force report. *Value Health* 10, 326–335
15 Yildirim, O. (2016) Opportunities and challenges for drug development: public?private partnerships, adaptive designs and big data. *Front. Pharmacol.* 7, 461
16 Liberati, A. *et al.* (2009) The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration. *PLoS Med.* 6, e1000100
17 Zhu, Y. *et al.* (2015) Zodiac: a comprehensive depiction of genetic interactions in cancer by integrating TCGA data. *J. Natl. Cancer Inst.* 107, djv129
18 Iqbal, U. *et al.* (2016) Cancer-disease associations: a visualization and animation through medical big data. *Comput. Methods Programs Biomed.* 127, 44–51
19 Grasso, M.A. (2015) Using big data to evaluate the association between peridontal disease DNA rheumatoid arthritis. *AMIA Annu. Symp. Proc.* 2015, 589–593
20 Camargo, C.A. *et al.* (2016) Pharmacoepidemiological study of long-acting β-agonist/inhaled corticosteroid therapy and asthma mortality. *Clin. Drug Invest.* 36, 993–999
21 Keenan, T.D. *et al.* (2015) Associations between primary open angle glaucoma, Alzheimer's disease and vascular dementia: record linkage study. *Br. J. Ophthalmol.* 99, 524–527
22 Rapsomaniki, E. *et al.* (2014) Blood pressure and incidence of twelve cardiovascular diseases: lifetime risks, healthy life-years lost, and age-specific associations in 1.25 million people. *Lancet* 383, 1899–1911
23 Hanauer, D.A. and Ramakrishnan, N. (2014) Modeling temporal relationships in large scale clinical associations. *J. Am. Med. Inform. Assoc.* 20, 332–341
24 Schuler, A. *et al.* (2016) Discovering patient phenotypes using generalized low rank models. *Pac. Symp. Biocomput.* 21, 144–155
25 Lafeuille, M.H. *et al.* (2013) Real-world corticosteroid utilization patterns in patients with metastatic castration-resistant prostate cancer in 2 large US administrative claims databases. *Am. Health Drug Benefits* 6, 307–316
26 Pfefferle, K.J. *et al.* (2014) Risk factors for manipulation after total knee arthroplasty: a pooled electronic health record database study. *J. Arthroplasty* 29, 2036–2038
27 Kaelber, D.C. *et al.* (2012) Patient characteristics associated with venous thromboembolic events: a cohort study using pooled electronic health record data. *J. Am. Med. Inform. Assoc.* 19, 965–972
28 Castro, V.M. *et al.* (2015) Validation of electronic health records phenotyping of bipolar disorder and controls. *Am. J. Psychiatry* 172, 363–372
29 Tufail, A. *et al.* (2014) The neovascular age-related macular degeneration database: multicenter study of 92 976 ranibizumab injections: report 1: visual acuity. *Ophthalmology* 121, 1092–1101
30 Van Herk-Sukel, M.P.P. *et al.* (2013) Pulmonary embolism, myocardial infarction, and ischemic stroke in lung cancer patients: results from a longitudinal study. *Lung* 191, 501–509
31 Spanier, B. *et al.* (2013) Incidence and mortality of acute and chronic pancreatitis in the Netherlands: a nationwide record-linked cohort study for the years 1995–2005. *World J. Gastroenterol.* 19, 3018–3026
32 Barnado, A. *et al.* (2017) Developing electronic health record algorithms to accurately identify patients with systemic lupus erythematosus. *Arthritis Care Res.* 69, 687–693
33 Kalyanam, J. *et al.* (2017) Exploring trends of nonmedical use of prescription drugs and polydrug abuse in the Twittersphere using unsupervised machine learning. *Addict. Behav.* 65, 289–295
34 Patel, V.N. and Kaelber, D.C. (2014) Using aggregated, de-identified electronic health record data for multivariate pharmacosurveillance: a case study of azathioprine. *J. Biomed. Inf.* 52, 36–42
35 Burwen, D.R. *et al.* (2012) Surveillance for Guillain–Barré syndrome after influenza vaccination among the Medicare population, 2009–2010. *Am. J. Public Health* 102, 1921–1927
36 Liu, Y. *et al.* (2012) Using temporal patterns in medical records to discern adverse drug events from indications. *AMIA Jt. Summits Transl. Sci. Proc.* 2012, 47–56
37 Michel, M.C. *et al.* (2000) Effect of diabetes on lower urinary tract symptoms in patients with benign prostatic hyperplasia. *J. Urol.* 163, 1725–1729

Reviews • INFORMATICS

Please cite this article in press as: Singh, G. et al. Real world big data for clinical research and drug development, Drug Discov Today (2018), https://doi.org/10.1016/j.drudis.2017.12.002