ANALYTIC METHODS USED IN REAL WORLD DATA BASED BIOMEDICAL RESEARCH- A SCOPING REVIEW

by
Chenyu Li

A thesis submitted to Johns Hopkins University in conformity with the requirements for the degree of Master of Science

Baltimore, Maryland
August 2020

Abstract

Background and Objective:

Real-world data (RWD) is characterized as data derived from multiple sources associated with the process in real-world practice in a heterogeneous patient population. There is a growing interest in using Real-World Data and Real-World Evidence in biomedical research since RWE presents an opportunity to extend the research beyond the typical limits of academia. However, the traditional statistics methods used in RWD analysis may lead to bias and challenge the credibility of RWE. To document what analytics methods have been used in RWD analysis, we conducted a sampled methodological review of methods used in EHRs based biomedical research.

Methods:

We developed an article database to document literature characteristics and analytical methods. We took a random sample of articles for detailed review. The primary outcome was proportion of articles using RWD methods. Meta-regressions were utilized to examine trends in proportion changes over time.

Results:

Of 88 papers reviewed in detail, 7 (8.0%) used the recommended Real-World Method (RWM). The proportion (and 95% confidence interval) of publications reporting having used RWM, performed sensitivity analysis, and handled missing data problem in 2019 were 11% (0, 26%), 17% (0, 34%) , and 22% (3%, 41%), respectively. Results of the sensitivity analysis showed the proportion of use RWM increased 0.4% per year, although this slope was statistically equivalent to 0.

Conclusions: The proportion of the EHRs based studies handling missing data, using RWM, or performing sensitivity analysis is disappointingly low. Although regulator guidelines, books, and academic meetings have suggested during the study period methods should be used in RWD analysis, the proper analytic methods are inadequately used in the published studies.

Keywords:

Real-World Evidences, Electronic Health Records, Analytic Methods, Missing Data, Sensitivity Analysis

Primary reader and Preceptor: Harold. P. Lehmann M.D. Ph.D.

Secondary reader : Karen Robinson Ph.D

Acknowledgement

I would like to express my sincere appreciation to my mentor, Professor Harold Lehmann, who guided me on my first thesis, always encouraged me in the past two years. Dr. Lehmann is one of the most knowledgable people I've met; he is my primary source to ask any questions. Without his persistent help, the goal of this thesis would not have been achieved.

I would like to thank my second reader Dr. Karen Robison, to give me valuable guidance at the beginning of this study and quickly provide me professional feedback on the thesis draft and revisions. I would like to show my gratitude to Ms.Claire Twose, who helped me define the search terms and search strategy for this thesis, And many thanks to Dr. Ray Alsheikh, who help with rating the articles and validating the dataset, providing insightful discussions about the research.

In addition, I am extremely grateful to my parents for supporting my education, loving me, and encouraging me throughout my life.

Table of Contents

List of Tables

List of Figures

Introduction

Real-World Data and Real-World Evidence in biomedical research

Using Real-World Data (RWD) to generate Real-World Evidence (RWE) is playing an increasing role in health care decisions worldwide [1] . There is a growing interest in using RWD in biomedical research by stakeholders, including policymakers, biomedical researchers, clinicians, and medical product developers. [2-8] Investigators believe that data objectively collected from a broad spectrum of therapeutic areas in routine care reflects the real-world practice. RWE that generated with RWD has the potential to support the regulatory decision-making, therapies discovery and evaluation, and clinical practice. The expected benefits of collecting data and extracting it from routine care settings are not only to improve study generalizability and reduce costs, but also to extend the available evidence for patients with substantial heterogeneity, multi-morbidity, and more severe forms of disease than would typically be allowed in a Random Controlled Trial (RCT) which is still the gold-standard of clinical research[9]. Using real-world data and real-world evidence in biomedical research will improve research feasibility and close the gap between clinical science and practice.[8]

Why now

In the United States, the U.S. Food & Drug Administration (FDA) has long been interested in using data generated in the real world to learn about medical products, including drugs, vaccines, biologics, and medical devices. In May 2008, the FDA launched the Sentinel Initiative, which is the national electronic system for researchers to monitor the safety of FDA-regulated medical products to protect public health [10]. Data from real-world practice were broadly collected after the Health Information Technology for Economic and Clinical Health Act (HITECH Act) was

signed in law in February 2009. [11]. Furthermore, The 21st Century Cures Act (Cures Act) in the U.S. was signed into law in 2016. The Congress requires *"Not later than 2 years after the date of enactment of the 21st Century Cures Act, the Secretary shall establish a program to evaluate the potential use of real-world evidence."* [12] The Act was designed to accelerate the development of medical products and to bring innovations and advancements to stakeholders who need them more efficiently and effectively[12 13]. It would also bring Congressional pressure on FDA inspectors to rely on RWD and RWE. FDA has developed guidelines on the various uses of RWE, for example, Best Practices for Conducting and Reporting Pharmacoepidemiological Safety Studies Using Electronic Health Records[14], Use of Electronic Health Record Data in Clinical Investigations-Guidance for Industry[15]. FDA's guidelines approved different research designs that can generate RWE, including but not limited to randomized trials, including big, simple trials, pragmatic trials, and observational studies. The guidelines of data analysis and RWE generation methodology are still under discussion. [1 14-17]

Why is it important

Although Real-World Data can be used in broad topics in biomedical research for multiple purposes, without a valid methodologic approach, including controlling source-data quality, choosing a proper study design, using correct analytical methods, sensitivity analysis of the results, RWD can lead to biased conclusions that cannot be used as the evidence to guide health care decision-making [17-19].

In the "Big Data Era", investigators are zealous about applying Artificial Intelligence, Machine Learning methods in the healthcare industry. Although the healthcare field has "big" volume and "big" variety data, the high-quality data that could be used to extract information and generate

clinical evidence using Artificial Intelligent (AI) or Machine Learning (ML) methods are limited. Using an improper or limited method to create Real-World Evidence from Real-World Data and then applying that RWE in real-world practice, may result in false treatment of patients, in waste of R&D funds, and in a delay of the study, each of which is the opposite of the expectation of using RWD in the research. Biomedical research results that were developed through machine learning methods, even with high accuracy, are not explainable and lack of robust causal reasoning[20]. Considering the cost and benefit of bringing new drugs and therapies into the biomedical research and development process, appropriate analytic methods designed for RWD and RWE should be used in the research.

Rationale for review

Several reviews have been done for the definition, opportunities, and challenges of using RWD and RWE, regulators and research institutions provided recommendations for how to generate RWE from RWD. EHRs are being used for research purposes, but because they comprise RWD, EHRs based research should use methods that suggested in books and guidelines that fit for RWD analysis. A research gap in systematically assess the published RWE quality exists. In conclusion, there is a need to review what analytic methods are used in biomedical research based on EMR data. In this paper, we review and document what data analysis methods are used in Electronic Health Records (EHRs) based biomedical research in the last ten years (2010 - 2019) .

Background

Real-World Data to Real-World Evidence

Real-world data (RWD) is characterized as data derived from multiple sources associated with process in real-world practice in a heterogeneous patient population. U.S. FDA defines Real-world data as "the data relating to patient health status and the delivery of health care routinely collected from a variety of sources." Example data sources include: 1) Electronic Health Records (EHRs) ; 2) Claims and billing records; 3) Product and device registries; 4) Patient-generated data including in home-use settings; 5) Data gathered from other sources that can inform on health status, such as mobile devices [1].

Real-World Evidence is a concept widely discussed in the Evidence-Based Medicine (EBM), but there is no one universally accepted definition for the RWE. U.S FDA defines RWE as *"the clinical evidence regarding the usage, and potential benefits or risks, of a medical product derived from analysis of real-world data"* [1]. The U.S. Congress defined the RWE as *"Any data on the application, or potential benefits or risks, of a product obtained from sources other than randomized clinical trials."* The Cures Act recognized the potential use of real-world evidence to help to support the approval of new indication for a drug, and to help to support a satisfied post-approval study requirement [12]. European Medicine Agency (EMA) *"Real-World Evidence meaning evidence coming from registries, electronic health records, and insurance data, etc. where studies may be required by regulators through scientific advice, CHMP or PRAC and the subsequent results are used to inform regulatory and potentially HTA decision-making "* [21]

Real-World Evidence is always compared with the gold-standard evidence from biomedical

research of Randomized Control Trial (RCT). Although RCT is the gold standard for establishing causal relationships analysis under ideal conditions, which includes a rigorous patient selection process and well-defined inclusion exclusion criteria, an RCT is not always practical. For instance, RCTs do not always represent the heterogeneous clinical population of patients, in practice. RWE offers insights into patient experiences in real-life environments, as opposed to the carefully planned conditions of experimental settings in healthcare. Also, RWD can be used to generate evidence complementing existing knowledge for the use of medical products in patients who are under-represented or excluded from the trial populations. Investigators expect RWE can provide information and expertise to researchers to answer questions in healthcare outcomes research, patient care, safety surveillance, and therapeutic development more effectively[9]. Based on the analysis of the 'real-world' medical history of patients dating back several years, data gathered from a larger patient population and evidence of real-practice patient compliance. RWE complements traditional RCT data and, as such, paints a broader picture of approaches used in the prevention, diagnosis, and management of particular diseases and long-term health. The application of RWD is not limited to RWE generation. Real-world data can be used to aid in the design of a clinical trial by: 1) assisting in the selection of research sites that are most likely to recruit test participants[22 23] , 2) providing a basis for power calculation[22 24] 3). creating a prior for a Bayesian statistical analysis, 4). and providing an alternative control group and guidance enrichment. (Jarow, LaVange et al. 2017, Sturmer, Wang et al. 2020). RWD can also be used during a trial to minimize duplication of data input, such as the medical history of the subject, automatic recording of adverse effects, and endpoints. [22]

A study listed 22 drugs, in which the FDA and EMA used RWE to support regulatory efficacy decisions leading to accelerated approval, full authorization or expansion of labels in the last 10 years [3]. While the expectation of using RWD and RWE is increasing, going from Real-world Data to Real-World Evidence is hard. Research in 2017 showed that the exist RWD can replicate only 15% RCTs in 2017 because of the data quality and study design limitation. [2 25] The results generated from RWD limited by unmeasured biases and confounding [20] [3] Therefore, Development of novel methodologies to produce RWE that provides adequate scientific evidence is needed.

Guidelines for Using RWE and RWD

Many guidelines and Act were published in last 10 years may facilitate the research on RWD based biomedical study. For instance, Best Practices for Conducting and Reporting Pharmacoepidemiological Safety Studies Using Electronic Health Records was published in [14], 21st Century Cures Act [12] and Guidelines for good pharmacoepidemiology practice (GPP) were published in 2016 [26].

Table 1 Guidelines for Using RWE and RWD

Name	**Published time**	**Agency**
Framework for FDA's Real-World Evidence Program[17]	2018	FDA
21st Century Cures Act[12]	2016	The Congress
Submitting Documents Using Real-World Data and Real-World Evidence to FDA for Drugs and Biologics-Guidance for Industry[16]	2019	FDA
Guidelines for good pharmacoepidemiology practice (GPP)[26]	2016	International Society of Pharmacoepidemiology
Use of Electronic Health Record Data in Clinical Investigations-Guidance for Industry[15]	July 2018	FDA
Best Practices for Conducting and Reporting Pharmacoepidemiological Safety Studies Using Electronic Health Records[14]	May 2013	FDA

Electronic Health Records as the Data source

Electronic Health Records (EHRs) are a major source of data that could generate RWD [17]. The scope of Electronic Health Records may be different in different health systems. We define the EHRs using the U.S. FDA definition in the guideline Use of Electronic Health Record Data in Clinical Investigations: “an EHR is an individual patient record contained within the EHR system. A typical individual EHR may include a patient’s medical history, diagnoses, treatment plans, immunization dates, allergies, radiology images, pharmacy records, and laboratory and test results” in this study [15]. The EHR is designed to optimize diagnosis or clinical care, as well as used to enhance the relevance of biomedical research. [17 19] Data obtained from EHRs always include structured information as the demographics, laboratory results(LONIC), procedures record (CTP code), clinical characteristics (ICD, LONIC), medications(patients outcomes, and unstructured information, for example, physician notes (notes from COPD system). It can also include unstructured data that, through text processing or natural language processing (NLP), can be turned into structured data as well.[27]

The HITECH Act encouraged health care providers to adopt Electronic Health Records Systems and improve health care data privacy and safety protection [11]. By the mid 2010s, over 86% of hospitals had their data stored in such records.[28] There is, therfore, a vast amount of healthcare data potentially available for study But how best to use such data is still under debate. The FDA is aware of recent attempts to use robust design and statistical methods to reproduce randomized study outcomes with observational studies and to derive general rules that could improve the likelihood of achieving reliable results using RWD in the design of observational studies[17].

Analytic Methods should be used for EHR data

EHRs comprise intrinsically longitudinal data that are collected in the routine delivery of patient care.[19] Challenges of using EHRs in biomedical research are recognized by many researchers. They can be categorized as IT systems challenges, data challenges, analytics methods challenges, and clinical knowledge challenges[19 25 29 30].

Table 2 Challenges of Using EHRs

Category	Problem
System challenge	Lack of standardization, e.g. data transfer protocols Interoperability within health information systems
Data challenge	Missing data / field Data quality validation Data gathering and integration Data storage and knowledge sharing Data publishing
Analytics	Study design based on research topic Study design to address missing data/ data problem Need for computable phenotype
Clinical knowledge	Are the results applicable for the clinical setting Cohort definition validation

As is stated in Guidelines for Good Pharmacoepidemiology Practices (GPP):

> *"Data analysis comprises comparisons and methods for analyzing and presenting results, categorizations, and procedures to control sources of bias and their influence on outcomes, e.g., the possible impact of biases due to selection bias, misclassification, confounding, and missing data.*[26]"

Furthermore, that the analysis method should

> *"be directed toward the unbiased estimation of the epidemiologic parameters of interest. The precision of effect estimates should be quantified using confidence intervals.*

Comparability of populations for pooled estimates should be assured, and missing important variables should be addressed. Interpretation of statistical measures, including confidence intervals, should be tempered with appropriate judgment and acknowledgements of potential sources of error and limitations of the analysis and should never be taken as the sole or rigid basis for concluding that there is or is not a relationship between an exposure and outcome. Sensitivity analyses should be conducted to examine the effect of varying potentially critical assumptions of the analysis".

"'*Any sensitivity analyses should be described.*'[26]"

From the guideline - Best Practices for Conducting and Reporting Pharmacoepidemiologic Safety Studies Using Electronic Healthcare Data- at a minimum, a research program should provide information on statistical models and tests, estimation of sample size, study meaning level ad strength, handling of missing values, analysis of subgroups and assessment of effect change, and confounding adjustment process [14].

As the Framework for FDA'S Real-World Evidence Program stated:

In considering whether data gathered through observational study designs are appropriate to generate RWE for the purpose of supporting effectiveness determinations, FDA intends to evaluate multiple questions of interest that could affect the ability to draw a reliable causal inference.[17]

Methods that could draw causal inference should be used in the RWD analysis.[17].

Rationale for using a sampled scoping review

Previous reviews and guidelines sought to understand the RWD and RWE at a theoretical level to assess the potential challenges and opportunities of using RWE. Real-world data analytic methods were discussed in industry guidelines, books, and academic meetings in past years.

However, it is unknown whether the analytic methods used in RWD analysis were applied as recommended. For these reasons, we conducted a scoping review to document analytic methods used in published RWD studies in the last 10 years, to identify any gaps in the RWD research. We followed the Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Review(PRISMA-ScR) checklist- an evidence-based minimum set of items for reporting scoping reviews- in this scoping review. [31]

Objectives and focus of review

To document what analytic methods investigators used from RWD to generate RWE. Focus on biomedical research papers that used Electronic Health Records as the main data source.

Methods

To document the current status of Real-World Evidence analytical methods in studies based on EMR data, we conducted a sampled methodological review to analyze the analytical methods used in the EHR-based biomedical research. From regulator guidelines, books, and RWD meeting recommendations we identified a list of analytic methodologies that should be used in Real-World Data analysis and coded them as "RWE methods". [14-20 26 32] In past 10 years, thousands of papers related to Real-world Evidence were published. To review the overall quality of the RWE papers, we used statistical inference to estimate the proportion of the papers used RWE methods out of all the papers. Two reviewers (CL and RA) checked the methods independently.

Scoping of the literature

This review aimed for summarizing analytic methods used in biomedical research that used EHRs as main data source from 2010-2019. We were looking for original quantitative research articles that analyzed data collected from real-world practice to answer biomedical questions written in English. The main concepts are "Electronic Health Records", "Biomedical Research"," Original Study", and time frame "2010-2019".

Search Strategy

We searched peer-reviewed articles in PubMed (MEDLINE) the major biomedical literatures database. The search term "Electronic Health Records" was extracted from MEDLINE / PubMed Search Strategy & Electronic Health Record Information Resources the version reviewed on May 24, 2019. [33], "Biomedical Research" was extracted from PubMed publication type "Study

Characteristics" which used as a broad strategy for research that use empirical methods include most of quantitative and qualitative biomedical research. Clinical Study[Publication Type] and Observational Study[publication type] is a subset of the Study Characteristics[Publication Type]. We exclude review articles by using "NOT Review[Publication Type] AND Systematic Review[Publication Type].[34] To limit the result to quantitative biomedical studies have data analytic methods, we added keywords "data"[All Fields] AND "analy*"[All Fields] to the search. We used PubMed clinical filters to focus on diagnosis, etiology, prognosis studies, the broad definition for searching diagnosis, etiology, and prognosis has sensitivity of 90% , 93% , and 90% , respectively.[35] Since the Electronic Health Systems were we limit the publication date to 2010/01/01-2019/12/31 using PubMed filters "2010/01/01" [PDat] : "2019/12/31" [PDat]. The detailed search strategy sees Appendix 1.

Developing of the Environment

EndNote

The articles extracted from PubMed Search were saved as a .nbib file (Appendix 2)mported to EndNote X9 library for further reading and annotation. Research objectives, study design, data source, analytic tools, analytic methods, and other relevant literature were highlighted in each article we read for this review. The notes were saved in the EndNote library for future evaluation and reproduce the reading process. (Appendix 3)

Microsoft Excel Articles Database

To manage the literature, to record the methods related to each literature, and to synthesize the evidence from the literature, we designed a relational database (Appendix 4) for the RWD articles we identified from the PubMed search. The Entity-Relationship Diagram is shown

Figure 1; detailed definitions were are available in the appendix (Table 9). The tables, Article_Review and Methods_Used_in_Literatures, were updated while Reviewer A (CL) reading. The database was designed, created and maintained by CL.

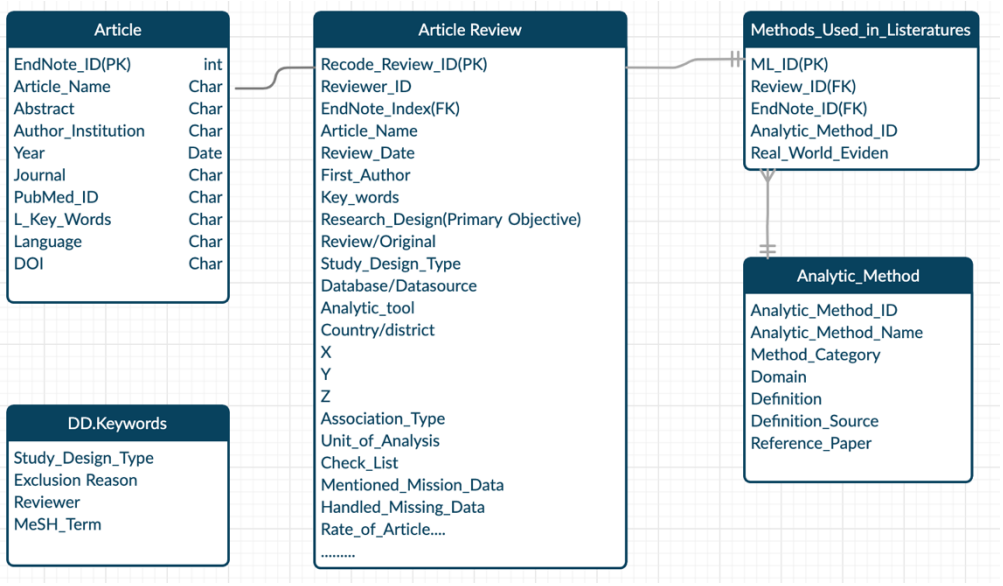


Figure 1 ER Diagram for managing the scoping review

Eligibility

This paper intends to include studies with original, quantitative, biomedical research used EHRs as the primary data source, written in English. The list of inclusion and exclusion criteria was developed based on two principles: (1) ensuring that the analysis in the paper was indeed of EHR-based data and (2) being wary of excluding studies likely to use RWE methods.

Inclusion criteria

We created four inclusion criteria (see Table 3 Inclusion Criteria No.1-4) to sensitively detect the articles we need. Based on this, we created a search strategy.

In the full-text reading process, we specified the inclusion criteria list, supplemented criteria No.3a,3b,3c, and No.5-7.

Table 3 Inclusion Criteria

Exclusion criteria

We grew the exclusion list in the course of the study. The primary RWD source in this review is routinely collected Electronic Health Record. Therefore we excluded the research that use Claims data, Genomic Data, Manually collected Registry Data, RCT data. Articles that focused on Physician behavior, Information System evaluation, health services evaluation, and new IT in healthcare were excluded. Also, the research used unstructured and semi-structured data which need Natural Language Processing or Text mining were excluded. Finally, we got 14 exclusion criteria and excluded ineligible papers while full-text reading.

Ideally, we should say something about having "grown this list in the course of the study, always with the guideline of (1) ensuring that the analysis in the paper was indeed of EHR-based data and (2) being wary of excluding studies likely to use RWE methods."

Number	Criterium
1	Original Data
2	Quantitative Study
3	Using the EHRs as the main source of data for analysis Allow collections of EHR data Either variables or outcomes should come from EHR Allow other source of data combined with the EHRs
4	Published year 2010-2019
5	Main Article written in English
6	Focus is on a biomedical question
7	National Data bank, if derived mostly from EHR

Analysis and Synthesis process

Thousands of literatures related to RWD and RWE were published, we aim to identify the proportion of papers used particular methods of the whol publication set. A random sampled literature set with a satisfactory sample size could present the whole literatures. We recorded

details in study type, study design and methods used in the article of a random sample of studies extractd from the MEDLINE.(RWE list see Appendix 4 Excel Database)

The key outcome variable was whether RWE methods were used. We documented the methods used in included papers, then matched the Real-World Methods (RWE) list we identified from regulator guidelines, books, and RWD meeting recommendations.[14 15 17 18 20 32 36] Any machine learning methods combined with causal inference also were considered as RWM. [32]

The missing data analytics process performed 1.Deletion methods with examining the sensitivity of results to the MCAR and MAR assumptions; [20] 2.Single impulation methods; 3. Model based methods , were tagged as handled missing data [20 36]

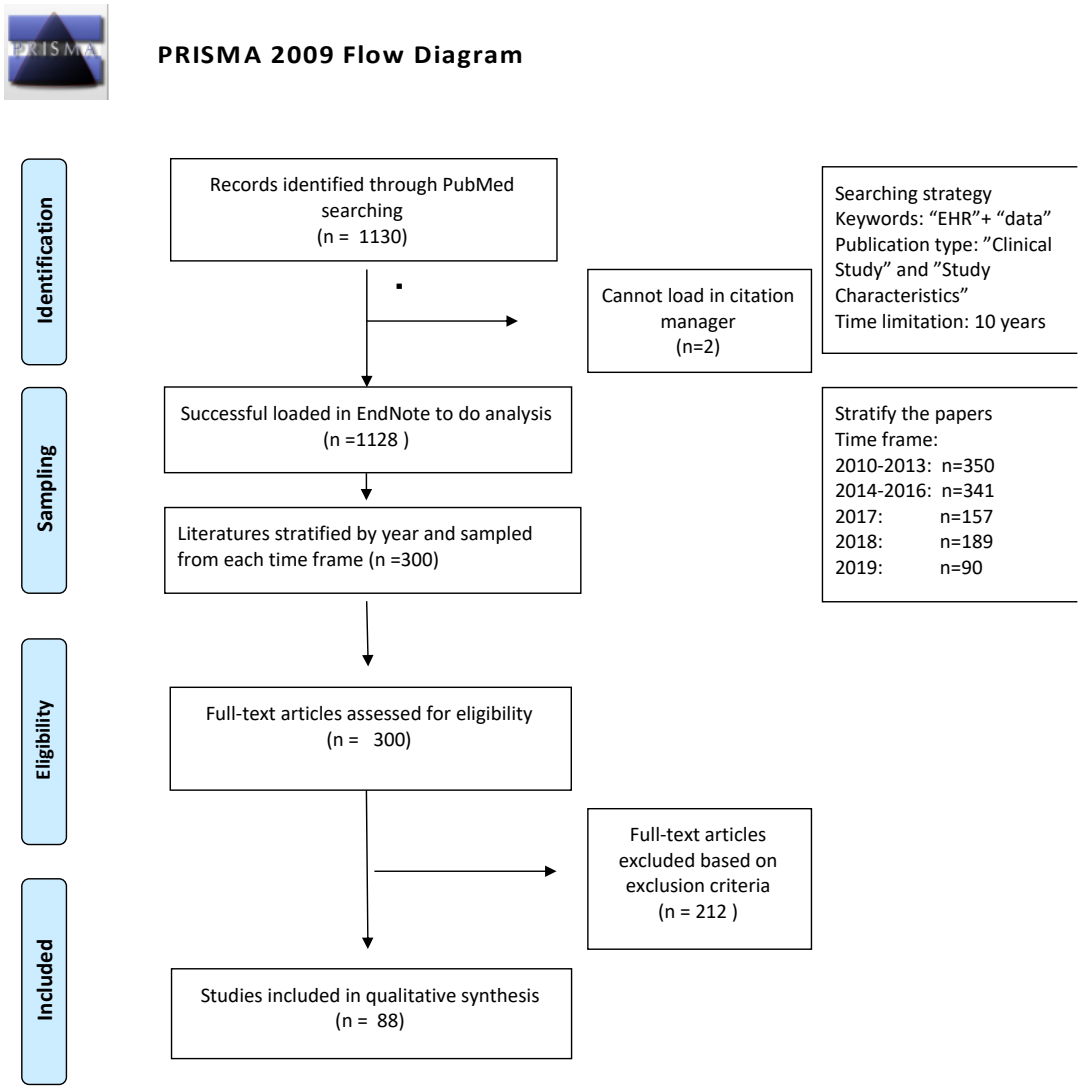
Every included paper was reviewed by 2 readers (CL and RA), and judgments logged. Characteristics of the study design type, Country/District, mentioned missing data, etc. were documented by CL after reading full-text. The second reader RA reviewed a random sample of the articles to recheck for errors in data documentdation or interpretation. Differences of opinion were discussed between the 2 readers and, if necessary, with the mentor (HL). Attention was paid to separate sensitivity analysis, which method was suggested be used whether RWE or more traditional methods are used [17 20 26] and missing data, which again is a concern in either framework[15 17 20 26 32].

The proportion of papers within each epoch using RWE, sensitivity analysis, or missing-data methods (of any sort) were calculated, along with the confidence interval of every such proportion (using bin size as the sample size), and graphed over time. A meta-regression across time was performed as well.

Results

Study Selection Flow

The study selection flow was summarized using PRISMA 2009 flow diagram (Figure 2)[37]



From: Moher D, Liberati A, Tetzlaff J, Altman DG, The PRISMA Group (2009). *P*referred *R*eporting *I*tems for *S*ystematic Reviews and *M*eta-*A*nalyses: The PRISMA Statement. PLoS Med 6(7): e1000097. doi:10.1371/journal.pmed1000097

Figure 2 PRISMA Flow Diagram

We conducted the literature searching on March 23rd, 2020 used the search strategy for MEDLINE described in Appendix 1. The final search results were exported into EndNote. Research papers were identified from the PubMed, the published paper number in each year has a trend of increase showed in the Figure 3 Extracted Papers by Year. According to the sample size calculation, for an α set at 0.05 , confidence interval at 5, n=1128, a total sample size of 287 will be needed.

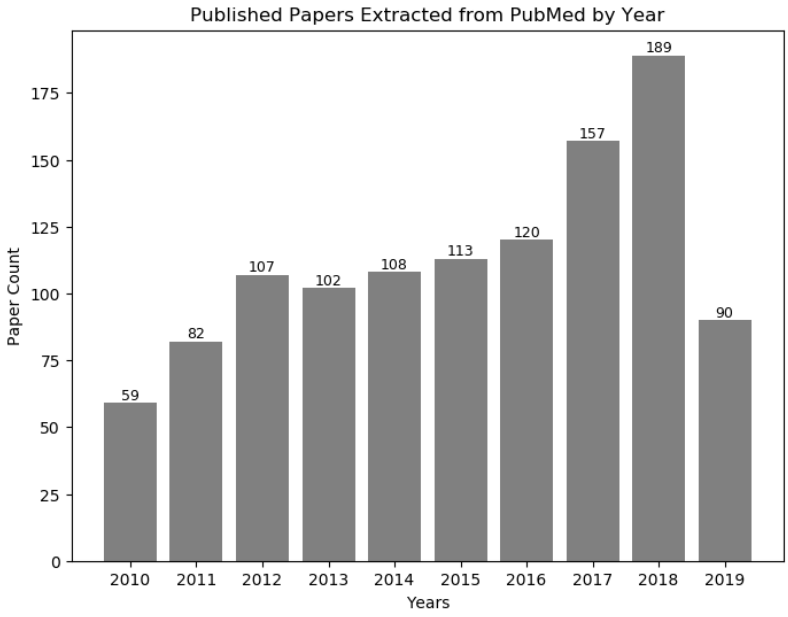


Figure 3 Extracted Papers by Year

Of the 1128 papers retrieved from PubMed, we sampled 300 articles, which is higher than the sample size needed to document the study details. We set 10 years as 5 unequal length epochs for further analysis. Randomly sampled 300 papers based on 5 epochs to build the Article_Review table. The sampling process was performed in Jupyter Notebook (Python 3.7 Random Module) for detailsfor details see Appendix 5.

Table 4 Articles Numbers by Epoch

Year	2010-2013	2014-2016	2017	2018	2019
Total papers	350	341	157	189	91
Proportion of total papers	31.0%	30.2%	13.9%	16.8%	8.1%
Sampled for full-text reading	100	50	50	50	50

After the full-text reading, and applying the inclusion/exclusion criteria. Exclude 212 papers reviewed by two readers based on the exclusion criteria list that generated from reading process. Excluded paper numbers with reasons see

Table 5. Finally, we ended with 88 papers in this review for detailed analytic methods evaluation.(Figure 4)

	Exclusion Reasons	Exclude number
1	Claim data only	8
2	EMR data only used to identify the cohort	7
3	Genomic data	2
4	Methodology papers	11
5	Not English	6
6	Patient generated health data only	6
7	Physician behavior, system evaluation, health services research [I.e., not biomedical]	32
8	Qualitative data only	5
9	Questionnaire/survey only	28
10	RCT data (data where a human being has abstracted the data [adds data quality; avoid curated data])	10
11	Registry (data where a human being has abstracted the data [adds data quality; avoid curated data])	31
12	Review papers	3
13	Technology question(Database build, data collection, datatransmission, IT infrastruction)	60
14	Text mining / NLP	3

Table 5 Exclusion Reasons

	Exclusion Reasons	Exclude number
1	Claim data only	8
2	EMR data only used to identify the cohort	7
3	Genomic data	2
4	Methodology papers	11
5	Not English	6
6	Patient generated health data only	6
7	Physician behavior, system evaluation, health services research [I.e., not biomedical]	32
8	Qualitative data only	5
9	Questionnaire/survey only	28
10	RCT data (data where a human being has abstracted the data [adds data quality; avoid curated data])	10
11	Registry (data where a human being has abstracted the data [adds data quality; avoid curated data])	31
12	Review papers	3
13	Technology question(Database build, data collection, datatransmission, IT infrastruction)	60
14	Text mining / NLP	3

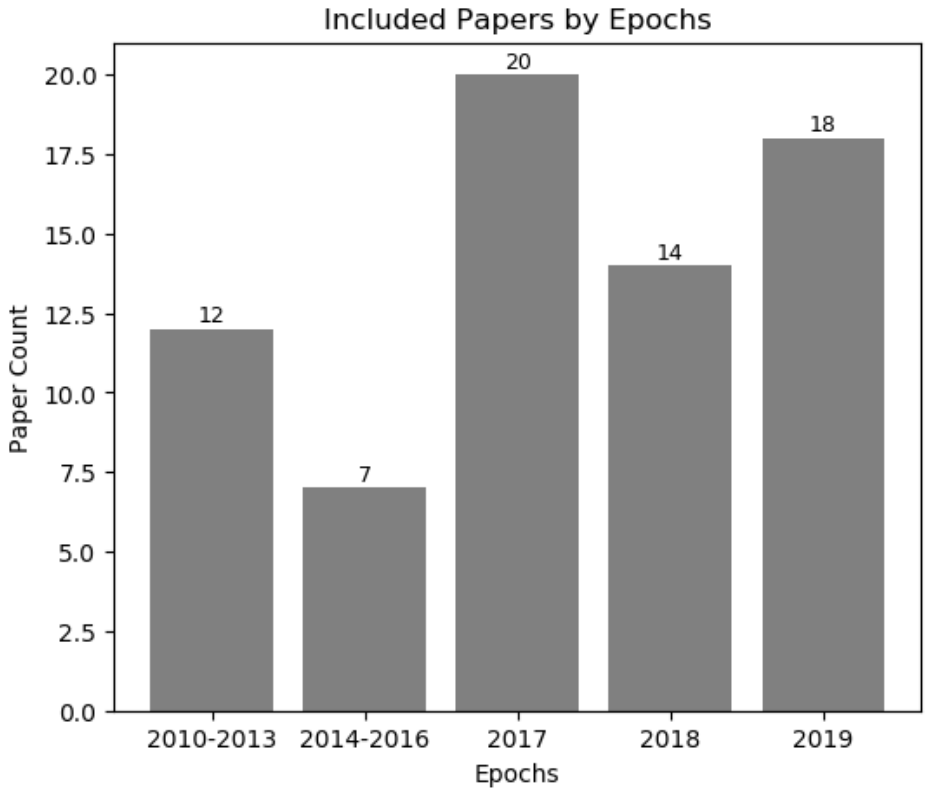


Figure 4 Included Papers by Epochs

Document Characteristics

The included paper characteristics were generated by Python TableOne package. We analyzed four characteristics by Epoch. Of the 88 included research articles, 67(77.9%) were designed as a retrospective cohort study, 6 (7%) were designed as a retrospective cross-sectional study. The missing recorded study design is 2 since the study design was not described in the article, and reviewers cannot identify the study type based on the scripts. Many articles have missing reported items recommended for the research report, only three papers stated their reports followed a checklist, and 15 (17.4%) did not report analytic tools used in the research. The Country/District was defined as the country in which the main population in the database. Of all the included research, 52 (59.3%) were conducted in the United States, 7 (8%) in the United Kingdom, and 6 (6.8%) in Korea.

Table 6 Included Paper Characteristics

INCLUDED PAPERS CHARACTERISTICS GROUPED BY EPOCHS								
		Missing	**Overall**	**2010-2013**	**2014-2016**	**2017**	**2018**	**2019**
N			88	26	12	18	14	18
STUDY_DESIGN_TYPE, N (%)	**retrospective cohort study**	2	67 (77.9)	21 (80.8)	10 (90.9)	12 (70.6)	10 (71.4)	14 (77.8)
	retrospective cross-sectional study		6 (7.0)	3 (11.5)		2 (11.8)		1 (5.6)
	cluster randomized pragmatic clinical trials		1 (1.2)	1 (3.8)				
	longitudinal, before/after study design		1 (1.2)			1 (5.9)		
	prospective cohort study		6 (7.0)			1 (5.9)	2 (14.3)	3 (16.7)
	quasi-experimental study		1 (1.2)			1 (5.9)		
	retrospective case–control study		2 (2.4)		1 (9.1)		1 (7.1)	
	retrospective chart review		1 (1.2)	1 (3.8)				
	proof of Concept Study		1 (1.2)				1 (7.1)	
COUNTRY/DISTRICT, N (%)	**USA**	0	52 (59.1)	16 (61.5)	6 (50.0)	11 (61.1)	7 (50.0)	12 (66.7)
	UK		7 (8.0)	2 (7.7)	2 (16.7)	3 (16.7)		
	French		2 (2.3)				1 (7.1)	1 (5.6)
	Brazil		1 (1.1)			1 (5.6)		
	Germany		2 (2.3)		1 (8.3)			1 (5.6)
	Italy		1 (1.1)	1 (3.8)				
	Japan		2 (2.3)			1 (5.6)		1 (5.6)
	Korea		6 (6.8)		1 (8.3)		4 (28.6)	1 (5.6)
	Netherland		4 (4.5)	2 (7.7)	1 (8.3)		1 (7.1)	
	Norway		1 (1.1)	1 (3.8)				
	Singapore		1 (1.1)					1 (5.6)
	South Korea		1 (1.1)	1 (3.8)				
	Spain		2 (2.3)	1 (3.8)		1 (5.6)		
	Sweden		1 (1.1)				1 (7.1)	

	Switzerland		1 (1.1)	1 (3.8)				
	Taiwan		1 (1.1)	1 (3.8)				
	Canada		2 (2.3)		1 (8.3)			1 (5.6)
	China		1 (1.1)			1 (5.6)		
MENTIONED_MISSION_DATA, N (%)	**No**	0	46 (52.3)	17 (65.4)	5 (41.7)	6 (33.3)	10 (71.4)	8 (44.4)
	Yes Data Analytic		9 (10.2)	1 (3.8)		3 (16.7)	2 (14.3)	3 (16.7)
	Yes Data Cleaning		14 (15.9)		3 (25.0)	6 (33.3)	2 (14.3)	3 (16.7)
	Yes Limitation		19 (21.6)	8 (30.8)	4 (33.3)	3 (16.7)		4 (22.2)
CHECK_LIST, N (%)	**Guidelines for good pharmacoepidemiology practices (GPP)**	85	1 (33.3)					1 (33.3)
	STROBE		2 (66.7)					2 (66.7)

Inter-rater reliability

The extent of agreement among data collectors was measure through interrater reliability testing. Reader A (CL) included 94 papers at first; after the second round review with reader B (RA), 88 papers were included. Cohen's Kappa was used to measure the eligibility process's inter-rater reliability, κ1=0.95, and the agreement between two reviewers was 98%.

The analytic methods were recorded in the Excel database by CL and reviewed by RA, the κ2 =0.97, with 99% agreement. Cohen suggested the Kappa result of 0.81–1.00 should be interpreted as an almost perfect agreement. [38]

$$\kappa = \frac{p_0 - p_e}{1 - p_e}$$

Equation 1 Cohen's Kappa

Proportion Analysis

The primary outcome of interest was the proportion of the Real-world Method used in RWD analysis. We used the sample (included papers) proportion to estimate the population (all published papers that meet the eligibility) proportion. The point estimator follows normal distribution, so the margin of error is the product of the Z value for the desired confidence level (in this case, we used Z=1.96 for 95% confidence) and the standard error of the point estimate. Confidence intervals(Equation 2 Confidence Interval Calculation) were calculated use Python (See, Appendix 5)

$$\hat{p} = \frac{X}{n}$$

$$\hat{\mathrm{p}} \sim \mathcal{N}\left(\frac{\mathrm{np}}{\mathrm{n}}, \frac{\sqrt{\mathrm{n\,p\,q}}}{\mathrm{n}}\right)$$

$$\text{Margin of error} = z_{\frac{\alpha}{2}}\left(\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right)$$

$$(lower\ bound, upper\ bound) = \left(\hat{p} - z_{\frac{\alpha}{2}}\left(\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right), \hat{p} + z_{\frac{\alpha}{2}}\left(\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right)\right)$$

Equation 2 Confidence Interval Calculation

The estimators of proportion were calculated through the equation above. As the results are shown in Table 7 , the estimated proportions of papers used Real-World Methods (RWM) in 2017, 2018, 2019 are 0.03,0.21,0.11, respectively. The 95% confidence interval of a proportion indicates a range within which, 95 out of 100 times, its estimated value will lie. [39]

Table 7 Proportion estimation and Confidence Interval

Estimated Proportion and Confidence Interval of Methods used in EHRs Based Research					
	2010-2013	**2014-2016**	**2017**	**2018**	**2019**
Real-World_Method	0.04 (0, 0.11)	0.08(0, 0.24)	0.03(0, 0.1)	0.21(0, 0.43)	0.11(0, 0.26)
Sensitivity_Analysis	0.19(0.04,0.34)	0.25(0.01, 0.5)	0.28(0.07, 0.48)	0.07(0, 0.21)	0.22(0.03, 0.41)
Handled_Missing_Data	0.12 (0, 0.24)	0.17(0, 0.38)	0.17(0, 0.34)	0.21(0, 0.43)	0.22(0.03, 0.41)

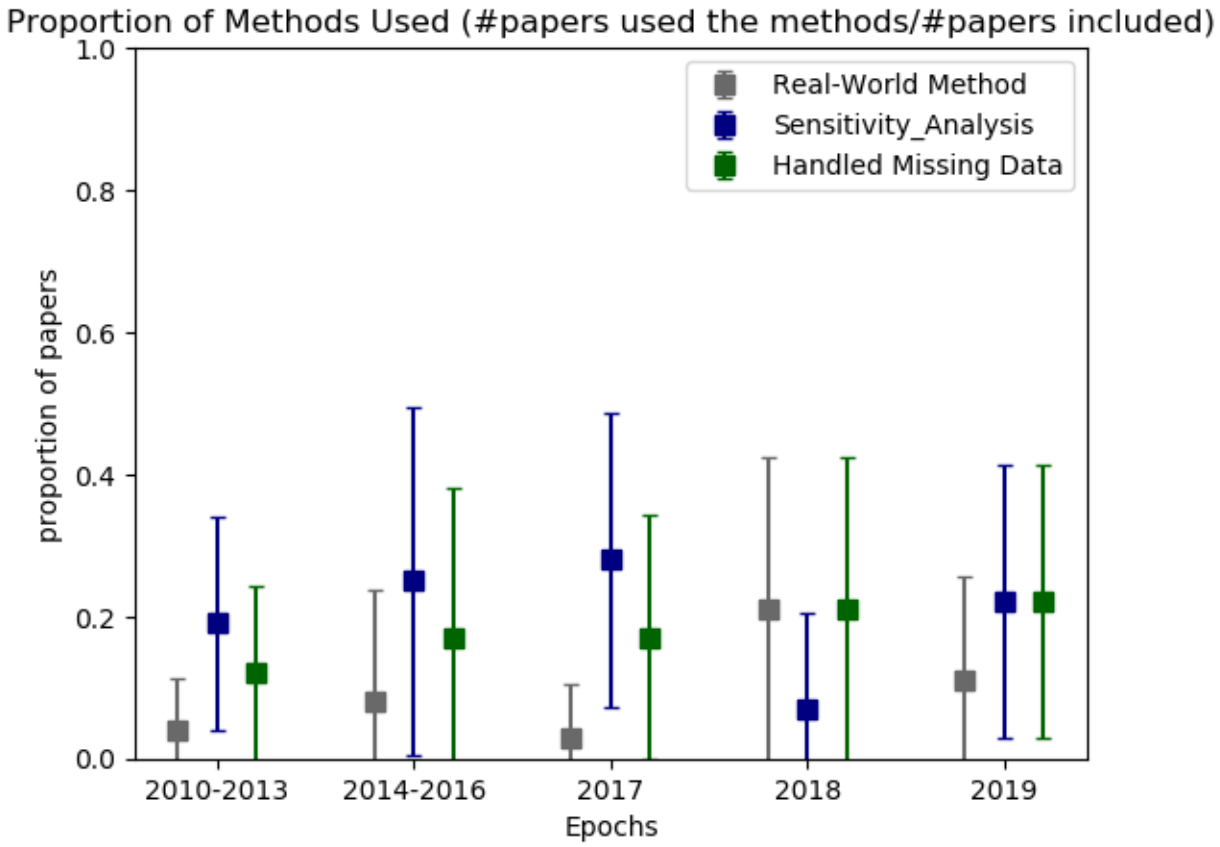


Figure 5 Proportion of Methods Used in the RWD Resesarch

As the proportion estimator is shown, 95% confidence interval shown in the Table 7 and Figure 5 the proportion of using proper methods in Real-World Evidence analysis is disappointingly low.

Meta-regressions

Meta-regression attempts to describe statistical variability in terms of study-level variables, thereby summarizing the information as a function rather than a single value. [40] The regression coefficient derived from a meta-regression analysis would explain how the outcome variable changes as the possible effect modifier with a cluster; in this case, proportion changes per

chronological time unit. The statistical significance of the coefficient is a test of if a linear relationship exists between the effect of action and the outcome variable.[40 41]

We conducted a mixed-effects meta-regression using restricted maximum-likelihood (ReML) using PyMARE, a package that does meta-analyses and meta-regressions in Python. [42]

Three meta-regression were done with time as the independent variable. Since the epoch length is unequal, we used the midpoint of each epoch in the analysis. As the result shown in Table 8. The proportion of use RWM increases by 0.4% per year; the upper bound is 3.4%, even use the small sample size and upper bound; the proportion of RWM used increases slowly. The p-values indicate that we cannot reject the null hypothesis : proportion of using RWM, Sensitivity Analysis or missing data methods does not change over year. The major measurements are proportion estimation, although the sample size is small, we do not believe that this may hinder the proportion estimation, confidence interval, and the conclusions of the review.

Table 8 Meta-regression for three methods

Mixed-effects Meta-regressions for three methods							
	name	estimate	Se	z-score	p-value	ci_0.025	ci_0.975
0	intercept	-0.0197	0.2301	0.0859	0.9314	-0.4709	0.43134
1	Real-World Methods	0.0049	0.0148	0.3310	0.7406	-0.0241	0.0340
	name	estimate	Se	z-score	p-value	ci_0.025	ci_0.975
0	intercept	17.9639	49.1902	0.3651	0.7149	-78.4473	114.3751
1	Sensitivity Analysis	-0.0088	0.0243	-0.3618	0.7174	-0.0566	0.0389
	name	estimate	Se	z-score	p-value	ci_0.025	ci_0.975
0	intercept	-23.1124	47.2377	0.4892	0.6246	115.6968	69.4719
1	Missing Data	0.0115	0.02344	0.4927	0.6221	-0.0343	0.0574

Discussion

Main findings

Multiple guidelines were published for how real-world data should be analyzed [14-17 26 32]. Our results suggest that, while the proportion of target studies using such methods have indeed risen over the past decade (EITHER rising from 10 to 20% OR meta-regression coefficient), even with the upper bound of the confidence interval, the actual rate is below 50% of studies. He estimate would be about 40%.

We believe our estimates are reliable and our conclusion, indeed, so. First, we took a novel sampling approach to the scoping review. Since our goal was a proportion of articles, such a sample approach is justified. Now, we did find, near the end of the study, that the NLM search strategy for Electronic Health Data includes several journal limitations, we conducted the new search on August 18th, 2020, 935 more records were extracted than the original searching, and the study needs to be extended to include the articles we missed. However, those articles constitute about 1/3 of all articles in each epoch. Even if 100% of such articles used the RWM we are seeking, the .

We biased our eligibility criteria to include studies that we would expect would use such methods. For instance, we did not exclude nine papers that used data from the National Trauma Data Bank (NTDB), [43-51] a large national database that attracts many researchers doing data analytic methodology related studies, the data were routinely collected and collated from trauma centers and trauma systems in the U.S. [52-54] However, the number of studies handled missing data is still low. A review in 2011 claims 10 % of articles used NTDB data handled

missing data [55], after almost 10 years, in our research 2 of those research handled missing data, and 4 mentioned missing data in limitation, only 1 used method suggested for Real-World Data. Third, we had 2 readers for judging inclusion/exclusion and determining the methods used within each included article. Our inter-rater reliabilities were 0.95 with agreement of 98% and 0.97 with agreement of 99%, both well above the kappa of .80 used as the practical threshold. [38] One of the major challenges in the analysis of EHRs is the missing data problem [19 56]. Forty-two included papers mentioned the missing data issue in the data cleaning or limitation session. However, only 15 of them handled the missing data. The estimated proportions of papers handled missing data problem are 0.17 (0,0.34), 0.21(0,0.43), and 0.02 (0.03,0.41) in 2017, 2018, and 2019, respectively. If the missingness is Missing at Completely at Random (MCAR) or Missing at Random (MAR), the probability of missing record is independent of observed data or outcome measurements, dropping the whole record with missing elements would not influence the estimator. However, many papers included in this review drop the missing records directly without giving proof of MCAR or MAR in a multivariate analysis. As a result, observations with missing values may lead to a biased result.

Real-world Methods we defined contains a list of methods that could analyze the causal effects of observed data, and machine learning methods with proper causal inference.[32] Causal inference is constrained by the assumptions made in the design and analysis of the research and this is especially evident when dealing with data on clinical health. [20] The proportion estimation of papers used RWM in 2018 is 21%, an upper bound of 43%, studies used RWM is disappointingly low. EHRs are observational data, from the EHRs the population being studied is the same that is being treated. The evidence generated without proper study design and analysis cannot be interpreted as meaningful information. Thus it is limited when to inform decision-

support. For example, only using linear regression on two variables extracted from EHR did not consider the counterfactual conditions; the result only can be construed as an association of two data variables. To better interpret RWD, investigators need the knowledge of informatics, epidemiology, and statistics is required.

Sensitivity Analysis seeks to determine the appropriateness of a particular analytic model and consider the impact of the model's conclusions. Sensitivity analysis should be performed after the analytic model was built to validate the study's primary results[20 26]. In our results, the proportion estimations of studies conducted sensitivity analysis are 0.28(0.07, 0.48), 0.07(0, 0.21), 0.22(0.03, 0.41) in 2017,2018 and 2019 , respectively. Although the guidelines proposed studies to do sensitivity analysis, only a small proportion of the study performed it.

This review found that proper methods designed for RWD were not correctly used in the published studies. To reduce biases in analytics, to enhance the cooperation of different background investigators, a standard process needs to be proposed and followed for the RWD results report.

Why might inadequate proper methods have been used and continue not to be used? At the first search, No facilitation/barrier study has been done, so we can only speculate on the following:

- Analysts of EHR data come from backgrounds with little exposure to EHR data;
- Informaticians who work with such data do not have the epidemiology and statistical background for their analysis;
- The tool supplied for these analyses (e.g., HADES), are not easily found, accessible, interoperable with standard models, or easilty reused.

Due to RWD's complexity, it is not accurate to use traditional data processing methods with large datasets. Despite the great value, EHRs may continue using inappropriate methods to

generate biased results against the original intention. In order to ensure internal and external validity in EHRs based research, researchers must determine whether the data are accurately extracted, adequately adjusted, correctly analyzed and cogently presented.[20] To understand and analyze the RWD in a proper method, it requires the investigators to collaborate in a multidisciplinary team that comprises clinicians, informaticians, epidemiologists, and biostatisticians (data scientists).

The OHDSI (The Observational Health Data Sciences and Informatics) developed tools to conduct real-world evidence generation.[30] From building Common Data Model (CDM), designing a study, defining cohort, building the analytics model, to generating the evidence, the RWD analytics is not a simple step. The set of tools are fantastic for conducting an observational study. However, for a small group of investigators, they may lack the ability to implement such a sophisticated toolset. There is a need to build an easily implemented research method decision-support toolset or standard RWE generation pipeline for existing Real-World Databases.

These suggestions have implications for education of statisticians and informaticians and for the need for statistical-analytic decision support tools, each of which is beyond the scope of this report.

Limitations

A limitation of this scoping review is it was not registered in any database. A search conducted in Cochrane Library and PROSPERO on August 18th, 2020, showed no similar systematic or scoping reviews were registered.

We have implemented a comprehensive search strategy, literature sampling , and synthesis process in accordance with the guidance for conducting methodological reviews. [57] The search strategy we used for the Electronic Health Record was retrieved form National Library of

Medicine MEDLINE / PubMed Search Strategy & Electronic Health Record Information Resources[33]. We found the strategy limited the literature to major journals in the EHRs research; the search strategy could lead to selection bias. For further investigation, more literature databases and adjusted search strategy need to be used.

The number of included papers in each epoch is small, and it should be increased for a more accurate analysis. However, we do not believe that this may hinder the proportion estimations and the conclusions of the review.

References

1. FDA. Real-World Evidence. 2020
2. Bartlett VL, Dhruva SS, Shah ND, Ryan P, Ross JS. Feasibility of Using Real-World Data to Replicate Clinical Trial Evidence. JAMA Netw Open 2019;**2**(10):e1912869 doi: 10.1001/jamanetworkopen.2019.12869[published Online First: Epub Date]|.
3. Baumfeld Andre E, Reynolds R, Caubel P, Azoulay L, Dreyer NA. Trial designs using real-world data: The changing landscape of the regulatory approval process. Pharmacoepidemiol Drug Saf 2019 doi: 10.1002/pds.4932[published Online First: Epub Date]|.
4. Blonde L, Khunti K, Harris SB, Meizinger C, Skolnik NS. Interpretation and Impact of Real-World Clinical Data for the Practicing Clinician. Adv Ther 2018;**35**(11):1763-74 doi: 10.1007/s12325-018-0805-y[published Online First: Epub Date]|.
5. Corrigan-Curay J, Sacks L, Woodcock J. Real-World Evidence and Real-World Data for Evaluating Drug Safety and Effectiveness. JAMA 2018;**320**(9):867-68 doi: 10.1001/jama.2018.10136[published Online First: Epub Date]|.
6. Franklin JM, Schneeweiss S. When and How Can Real World Data Analyses Substitute for Randomized Controlled Trials? Clin Pharmacol Ther 2017;**102**(6):924-33 doi: 10.1002/cpt.857[published Online First: Epub Date]|.
7. Makady A, de Boer A, Hillege H, Klungel O, Goettsch W. What Is Real-World Data? A Review of Definitions Based on Literature and Stakeholder Interviews. Value Health 2017;**20**(7):858-65 doi: 10.1016/j.jval.2017.03.008[published Online First: Epub Date]|.
8. Miksad RA, Abernethy AP. Harnessing the Power of Real-World Evidence (RWE): A Checklist to Ensure Regulatory-Grade Data Quality. Clin Pharmacol Ther 2018;**103**(2):202-05 doi: 10.1002/cpt.946[published Online First: Epub Date]|.
9. Sherman RE, Anderson SA, Dal Pan GJ, et al. Real-World Evidence - What Is It and What Can It Tell Us? N Engl J Med 2016;**375**(23):2293-97 doi: 10.1056/NEJMsb1609216[published Online First: Epub Date]|.
10. FDA. FDA's Sentinel Initiative. Secondary FDA's Sentinel Initiative 2008. https://www.fda.gov/safety/fdas-sentinel-initiative.
11. HITECH. PUBLIC LAW 111-5 HITECH Act 2009
12. Congress t. 21st Century Cures Act 2016
13. Kesselheim AS, Avorn J. New "21st Century Cures" Legislation: Speed and Ease vs Science. JAMA 2017;**317**(6):581-82 doi: 10.1001/jama.2016.20640[published Online First: Epub Date]|.
14. FDA. Best Practices for Conducting and Reporting Pharmacoepidemiologic Safety Studies Using Electronic Healthcare Data. 2013
15. FDA. Use of Electronic Health Record Data in Clinical Investigations Guidance for Industry. 2018
16. FDA. Submitting Documents Using Real-World Data and Real-World Evidence to FDA for Drugs and Biologics. Guidance for Industry. 2019
17. FDA. Framework for FDA's Real-World Evidence Program. 2018
18. Sturmer T, Wang T, Golightly YM, Keil A, Lund JL, Jonsson Funk M. Methodological considerations when analysing and interpreting real-world data. Rheumatology (Oxford)

2020;**59**(1):14-25 doi: 10.1093/rheumatology/kez320[published Online First: Epub Date]|.
19. Cowie MR, Blomster JI, Curtis LH, et al. Electronic health records to facilitate clinical research. Clin Res Cardiol 2017;**106**(1):1-9 doi: 10.1007/s00392-016-1025-6[published Online First: Epub Date]|.
20. MIT. Secondary Analysis of Electronic Health Records. 2016
21. EMA. Real World Evidence Data Collection. 2016
22. Jarow JP, LaVange L, Woodcock J. Multidimensional Evidence Generation and FDA Regulatory Decision Making: Defining and Using "Real-World" Data. JAMA 2017;**318**(8):703-04 doi: 10.1001/jama.2017.9991[published Online First: Epub Date]|.
23. Thorpe KE, Zwarenstein M, Oxman AD, et al. A pragmatic-explanatory continuum indicator summary (PRECIS): a tool to help trial designers. J Clin Epidemiol 2009;**62**(5):464-75 doi: 10.1016/j.jclinepi.2008.12.011[published Online First: Epub Date]|.
24. Swift B, Jain L, White C, et al. Innovation at the Intersection of Clinical Trials and Real-World Data Science to Advance Patient Care. Clin Transl Sci 2018;**11**(5):450-60 doi: 10.1111/cts.12559[published Online First: Epub Date]|.
25. Ramamoorthy A, Huang SM. What Does It Take to Transform Real-World Data Into Real-World Evidence? Clin Pharmacol Ther 2019;**106**(1):10-18 doi: 10.1002/cpt.1486[published Online First: Epub Date]|.
26. Public Policy Committee ISoP. Guidelines for good pharmacoepidemiology practice (GPP). Pharmacoepidemiol Drug Saf 2016;**25**(1):2-10 doi: 10.1002/pds.3891[published Online First: Epub Date]|.
27. Hong N, Wen A, Shen F, et al. Developing a scalable FHIR-based clinical data normalization pipeline for standardizing and integrating unstructured and structured electronic health record data. JAMIA Open 2019;**2**(4):570-79 doi: 10.1093/jamiaopen/ooz056[published Online First: Epub Date]|.
28. HealthIT. Office-based Physician Electronic Health Record Adoption. Secondary Office-based Physician Electronic Health Record Adoption 2019. https://dashboard.healthit.gov/quickstats/pages/physician-ehr-adoption-trends.php.
29. Eichler HG, Bloechl-Daum B, Broich K, et al. Data Rich, Information Poor: Can We Use Electronic Health Records to Create a Learning Healthcare System for Pharmaceuticals? Clin Pharmacol Ther 2019;**105**(4):912-22 doi: 10.1002/cpt.1226[published Online First: Epub Date]|.
30. Ryan. P DJ. The Book of OHDSI. The book of OHDSI -Observational Health Data Sciences and Informatics, 2020.
31. PRISMA Extension for Scoping Reviews (PRISMA-ScR): Checklist and Explanation. Annals of Internal Medicine 2018;**169**(7):467-73 doi: 10.7326/m18-0850 %m 30178033[published Online First: Epub Date]|.
32. FDA. Developing Real-World Data and Evidence to Support Regulatory Decision-Making, 2019.
33. PubMed. MEDLINE / PubMed Search Strategy & Electronic Health Record Information Resources. 2019
34. PubMed. Publication Characteristics (Publication Types) with Scope Notes. 2020
35. PubMed. PubMed Clinical Queries 2020
36. Holmes JH. Knowledge Discovery in Biomedical Data: Theory and Methods. Methods in Biomedical Informatics, 2014:179-240.

37. Moher D, Liberati A, Tetzlaff J, Altman DG, The PG. Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. PLOS Medicine 2009;**6**(7):e1000097 doi: 10.1371/journal.pmed.1000097[published Online First: Epub Date]|.
38. McHugh ML. Interrater reliability: the kappa statistic. Biochem Med (Zagreb) 2012;**22**(3):276-82
39. Thomas R, Braganza A, Oommen LM, Muliyil J. Confidence with confidence intervals. Indian J Ophthalmol 1997;**45**(2):119-23
40. Health CP. Meta-Regression. Secondary Meta-Regression. https://www.publichealth.columbia.edu/research/population-health-methods/meta-regression.
41. Cochrane Handbook for Systematic Reviews of Interventions-Meta-regression. Secondary Cochrane Handbook for Systematic Reviews of Interventions-Meta-regression 2020. https://handbook-5-1.cochrane.org/chapter_9/9_6_4_meta_regression.htm.
42. PyMARE: Python Meta-Analysis & Regression Engine [program], 2020.
43. Cook A, Osler T, Glance L, et al. Comparison of two prognostic models in trauma outcome. Br J Surg 2018;**105**(5):513-19 doi: 10.1002/bjs.10764[published Online First: Epub Date]|.
44. Sabat J, Hsu CH, Samra N, et al. Length of Stay and ICU Stay Are Increased With Repair of Traumatic Superior Mesenteric Vein Injury. J Surg Res 2019;**242**:94-99 doi: 10.1016/j.jss.2019.04.043[published Online First: Epub Date]|.
45. Bennett KS, DeWitt PE, Harlaar N, Bennett TD. Seizures in Children With Severe Traumatic Brain Injury. Pediatr Crit Care Med 2017;**18**(1):54-63 doi: 10.1097/pcc.0000000000000948[published Online First: Epub Date]|.
46. Patel MS, Malinoski DJ, Zhou L, Neal ML, Hoyt DB. Penetrating oesophageal injury: a contemporary analysis of the National Trauma Data Bank. Injury 2013;**44**(1):48-55 doi: 10.1016/j.injury.2011.11.015[published Online First: Epub Date]|.
47. Phillips B, Turco L, McDonald D, Mause E, Walters RW. A subgroup analysis of penetrating injuries to the pancreas: 777 patients from the National Trauma Data Bank, 2010-2014. J Surg Res 2018;**225**:131-41 doi: 10.1016/j.jss.2018.01.014[published Online First: Epub Date]|.
48. Haider AH, Hashmi ZG, Zafar SN, et al. Minority trauma patients tend to cluster at trauma centers with worse-than-expected mortality: can this phenomenon help explain racial disparities in trauma outcomes? Ann Surg 2013;**258**(4):572-9; discussion 79-81 doi: 10.1097/SLA.0b013e3182a50148[published Online First: Epub Date]|.
49. Strutt J, Flood A, Kharbanda AB. Shock Index as a Predictor of Morbidity and Mortality in Pediatric Trauma Patients. Pediatr Emerg Care 2019;**35**(2):132-37 doi: 10.1097/pec.0000000000001733[published Online First: Epub Date]|.
50. Fierro N, Inaba K, Aiolfi A, et al. Motocross versus motorcycle injury patterns: A retrospective National Trauma Databank analysis. J Trauma Acute Care Surg 2019;**87**(2):402-07 doi: 10.1097/ta.0000000000002355[published Online First: Epub Date]|.
51. Petersen S, Simms ER, Guidry C, Duchesne JC. Impact of hormonal protection in blunt and penetrating trauma: a retrospective analysis of the National Trauma Data Bank. Am Surg 2013;**79**(9):944-51

52. Moore L, Hanley JA, Lavoie A, Turgeon A. Evaluating the validity of multiple imputation for missing physiological data in the national trauma data bank. J Emerg Trauma Shock 2009;**2**(2):73-9 doi: 10.4103/0974-2700.44774[published Online First: Epub Date]|.
53. Haider AH, Hashmi ZG, Zafar SN, et al. Developing best practices to study trauma outcomes in large databases: an evidence-based approach to determine the best mortality risk adjustment model. J Trauma Acute Care Surg 2014;**76**(4):1061-9 doi: 10.1097/TA.0000000000000182[published Online First: Epub Date]|.
54. NTDB Data Dictionary 2019 Revision, 2019.
55. Haider AH, Saleem T, Leow JJ, et al. Influence of the National Trauma Data Bank on the study of trauma outcomes: is it time to set research best practices to further enhance its impact? J Am Coll Surg 2012;**214**(5):756-68 doi: 10.1016/j.jamcollsurg.2011.12.013[published Online First: Epub Date]|.
56. Bayley KB, Belnap T, Savitz L, Masica AL, Shah N, Fleming NS. Challenges in using electronic health record data for CER: experience of 4 learning organizations and solutions applied. Med Care 2013;**51**(8 Suppl 3):S80-6 doi: 10.1097/MLR.0b013e31829b1d48[published Online First: Epub Date]|.
57. Gentles SJ, Charles C, Nicholas DB, Ploeg J, McKibbon KA. Reviewing the research methods literature: principles and strategies illustrated by a systematic overview of sampling in qualitative research. Systematic Reviews 2016;**5**(1):172 doi: 10.1186/s13643-016-0343-0[published Online First: Epub Date]|.

Appendix

Appendix 1 Search Strategy Details (on March 23rd, 2020)

Keyworks	Details	Reference
Electronic Health Record (#1)	((health information exchange [tw] OR hie [tw] OR rhio [tw] OR regional health information organization [tw] OR hl7 [tw] ORhealth level seven [tw] OR unified medical language system [majr] OR umls [tw] OR loinc [tw] OR rxnorm [tw] OR snomed [tw] OR icd9 cm [ti] OR icd 9 cm [ti] OR icd10 [ti] OR icd 10 [ti] OR metathesaurus [tw] OR patient card [tw] OR patient cards [tw] OR health card [tw] OR health cards [tw] OR electronic health data [tw] OR personal health data [tw] OR personal health record [tw] OR personal health records [tw] OR Health Records, Personal [Majr] OR Health Record, Personal [Majr] OR ehealth [tw] OR e-health [tw] OR medical informatics application [mh] OR medical informatics applications [mh] OR medical records system, computerized [mh] OR medical records systems, computerized [mh] OR computerized patient medical records [tw] OR automated medical record system [tw] OR automated medical record systems [tw] OR automated medical records system [tw] OR automated medical records systems [tw] OR computerized medical record [tw] OR computerized medical records [tw] OR computerized patient records [tw] OR computerized patient record [tw] OR computerized patient medical record [tw] OR electronic health record [tw] OR electronic health records [tw] OR Electronic Health Record [Majr] OR Electronic Health Records [Majr] OR electronic patient record [tw] OR electronic patient records [tw] OR	MEDLINE / PubMed Search Strategy & Electronic Health Record Information Resources https://www.nlm.nih.gov/services/queries/ehr_details.html

	electronic medical record [tw] OR electronic medical records [tw] OR electronic healthcare records [tw] OR electronic healthcare record [tw] OR electronic health care record [tw] OR electronic health care records [tw] OR archives [majr] OR ehr [tw] OR ehrs [tw] OR phr [tw] OR phrs [tw] OR emr [tw] OR emrs [tw] OR Health Information Systems [Majr] OR health information interoperability[mh] OR health information interoperability[tw]) AND (medical record [ti] OR medical records [mh] OR medical records [ti] OR patient record [ti] OR patient records [ti] OR patient health record [ti] OR patient health records [ti] OR patient identification system [mh] OR patient identification systems [mh] OR Patient Outcome Assessment[Majr] OR Patient Discharge Summaries[Majr] OR healthcare record [ti] OR healthcare records [ti] OR health care record [ti] OR health care records [ti] OR health record [ti] OR health records [ti] OR hospital information system [tw] OR hospital information systems [tw] OR umae [ti] OR attitude to computers [mh] OR medical informatics [ti] OR Information Technology[mh] OR Information Technology[tw])) OR ((medical records systems, computerized [majr] OR medical records systems, computerized [mh] OR computerized patient medical record [tw] OR computerized patient medical records [tw] OR automated medical record system [tw] OR automated medical record systems [tw] OR automated medical records system [tw] OR automated medical records systems [tw] OR computerized medical record [tw] OR computerized medical records [tw] OR computerized patient records [tw] OR computerized patient record [tw] OR patient generated health data[mh] OR patient generated health data[tw] OR electronic health record [tw] OR	

	electronic health records [tw] OR electronic patient record [tw] OR electronic patient records [tw] OR electronic medical record [tw] OR electronic medical records [tw] OR electronic healthcare records [tw] OR electronic healthcare record [tw] OR electronic health care record [tw] OR electronic health care records [tw] OR unified medical language system [majr] OR unified medical language system [tw] OR umls [tw] OR loinc [tw] OR rxnorm [tw] OR snomed [tw] OR icd9 cm [ti] OR icd 9 cm [ti] OR icd10 [ti] OR icd 10 [ti] OR metathesaurus [tw] OR ehr [tw] OR ehrs [tw] OR phr [tw] OR phrs [tw] OR emr [tw] OR emrs [tw] OR meaningful use [tiab] OR meaningful use [tw] OR Meaningful Use [Majr]) AND (j ahima [ta] OR j am med inform assoc [ta] OR amia annu symp proc [ta] OR health data manag [ta] OR int j med inform [ta] OR yearb med inform [ta] OR telemed j e health [ta] OR stud health technol inform [ta])	
Biomedical Quantitative Study (#2)	"Study Characteristics"[Publication Type] AND “data”[All fileds] AND “analy*”[All Fields] NOT "Review"[Publication Type] NOT “Systematic Review"[Publication Type]	Publication Characteristics (Publication Types) with Scope Notes 2020 MeSH Pubtypes https://www.nlm.nih.gov/mesh/pubtypes.html
Clinical Filter (#3)	(sensitiv*[Title/Abstract] OR sensitivity and specificity[MeSH Terms] OR diagnose[Title/Abstract] OR diagnosed[Title/Abstract] OR diagnoses[Title/Abstract] OR diagnosing[Title/Abstract] OR diagnosis[Title/Abstract] OR diagnostic[Title/Abstract] OR diagnosis[MeSH:noexp] OR diagnostic * [MeSH:noexp] OR diagnosis,differential[MeSH:noexp] OR	Clinical Queries using Research Methodology Filters

	diagnosis[Subheading:noexp]) OR (risk*[Title/Abstract] OR risk*[MeSH:noexp] OR risk *[MeSH:noexp] OR cohort studies[MeSH Terms] OR group[Text Word] OR groups[Text Word] OR grouped [Text Word]) OR (incidence[MeSH:noexp] OR mortality[MeSH Terms] OR follow up studies[MeSH:noexp] OR prognos*[Text Word] OR predict*[Text Word] OR course*[Text Word])	https://www.ncbi.nlm.nih.gov/books/NBK3827/table/pubmedhelp.T.clinical_queries_using_rese/
From 2010/01/01-2019/12/31 (#4)	"2010/01/01"[PDat] : "2019/12/31"[PDat]	

Appendix 2 PubMed searched result
PubMed file -github

Appendix 3 EndNote Library
EndNote Library-github

Appendix 4 Excel Database
Excel-Database-github

Table 9 Database Filed Definitions

Article		Source of truth for article entities; data taken from EndNote
	EndNote_ID	From EndNote
	Article_Name	From EndNote
	Abstract	From EndNote
	Author_Institution	From EndNote
	Year	From EndNote
	Journal	From EndNote
	PubMed_ID	From EndNote
	L_Key_Words	From EndNote
	Language	From EndNote
	DOI	From EndNote
Article_Review		One row per review; allows multiple reviews per article
	Recode_Review_ID	Primary Key
	Reviewer_ID	DD.Keyworks List
	EndNote_Index	Foreign key for Article table
	Article_Name	vlookup from Article table
	Review_Date	Manually enter timestamp
	First_Author	Manually enter
	Key_words	Manually enter
	Research_Design(Primary Objective)	Manually enter
	Review/Original	Manually enter
	Study_Design_Type	Select from DD.Keywords_List Study Type
	Database/Datasource	Manually enter
	Analytic_tool	Manually enter

	Country/district	Manually enter
	X	Manually enter
	Y	Manually enter
	Z	Manually enter
	Association_Type	Manually enter
	Unit_of_Analysis	Manually enter
	Check_List	Manually enter
	Mentioned_Mission_Data	Manually enter
	Handled_Missing_Data	Manually enter
	Rate_of_Article	Manually enter
	Include_in_Research	Manually enter
	Exclusion Reason	Select from Exclusion Criteria(DD.Keywords_List)
	Real-World_Method	TRUE/FALSE searched from Methods_Used_In_Literature table
	Sensitivity_Analysis	TRUE/FALSE searched from Methods_Used_In_Literature table
	Other_Notes	
Methods_Used_in_Literatures		One row per analytic method; enables multiple methods per review
	ML_ID	Methods records ID
	Review_ID	foreign key for Article_Review table
	EndNote_ID	Foreign key for Article table
	Analytic_Method_ID	foreign key for DD.Analytic_Method table
	Real_World_Evidence	vlookup from DD.Analytic_Method table
Analytic_Method		
	Analytic_Method_ID	Primary Key
	Analytic_Method_Name	Manually enter
	Method_Category	Enter based on Guidelines
	Domain	Manually enter
	Definition	Manually enter
	Definition_Source	Manually enter
	Reference_Paper	Manually enter
DD.Keywords		
	Study_Design_Type	A list generated from reading process
	Exclusion Reason	A list defined before reading
	Reviewer	A list defined before reading
	MeSH_Term	A list extracted from EndNote

Appendix 5 Analytic Code-Python 3.7

Analytic Code-github

Biographical Statement

Chenyu Li

929 North Wolfe.St, Baltimore, Maryland 21205

Tel:(+1) 443-635-8987 chenyu.li@jhmi.edu linkedin.com/in/chenyu-li-80375196/

PROFILE

Master of Science candidate concentrating in biomedical informatics, interested in Real world healthcare data analysis and decision support. Bachelor of Management specialized in Information Systems Management.

EDUCATION BACKGROUND

- 08/2018-08/2020 Johns Hopkins University School of Medicine GPA:3.8

Master of Science in Health Sciences Informatics-Research
Certificate in Health Finance and Management

- 08/2015-05/2016 Illinois Institute of Technology GPA:4.0

Exchange Student - Information Technology Management

- 09/2013-06/2018 Beijing University of Posts and Telecommunications (BUPT) GPA:3.4

Bachelor of Management in Information Management and Information Systems

ACADEMIC PROJECTS

Johns Hopkins University *Research Student Assistant* Baltimore, MD 12/2018-

- EHR Data quality analysis project- analyzed simulated weight data in a large database using R, summarized data quality pattern.
- Cost-benefit analysis of a Telemedicine program in pediatric anesthesia preoperative-prepared program workflow, study design and data analysis plan.
- Graduate thesis: Analytic Methods Used in Real-World Data Based Biomedical Research- A Sampled Methodological Scoping Review

Mentor: Harold Lehmann

Clinic Data Analysis with Python Baltimore, MD 01/2020-03/2020

- Used Python libraries Pandas, Matplotlib, Seaborn, Sklearn, PySpark to do data cleaning, visualization, analysis on Hopkins Precision Medicine Platform Asthma dataset;
- Accomplished DataCamp certifications: Python Programmer, Data Analysis with Python, Data Science for everyone with Python

Real-time Disease Surveillance Systems for COPD Baltimore, MD 01/2019-03/2019

- Designed workflow and data-flow for a disease surveillance system;
- Visualized Chronic Obstructive Pulmonary Disease (COPD) related data using Tableau;
- Created a website for demonstrating the project.

General Hospital of the People's Liberation Army (PLAGH) Beijing, China 12/2017-05/2018

Medical Engineering Center, Research Assistant

- Conducted literature review on 3 Common Data Model used in Biomedical research, summarized cons and pros of CDMs and presented the results to clinicians;

- Collaborated in a clinical research data sharing system deployment;
- *Extracted* 10-year Emergency Department EHR data, *Transformed* into ODHSI OMOP CDM 5.0 version used SQL and PostgreSQL database, and *Loaded* into a clinical research data sharing system;
- Assisted Database performance monitoring and trouble shooting.
- Advisor: Zhengbo Zhang

INTERNSHIPS

American College of Radiology, Application Development Intern — Reston, VA — 05/2019-08/2019

- Developed Clinical Decision Support (CDS) tool through Agile methods;
- Recognized workflow, dataflow from research papers and converted 6 research papers into radiology Clinical Decision Support tool modules using XML;
- Analyzed XML schemas used in 2 different organizations, communicated with members in different groups, and created XML schema mapping rules.

Project Manager : Sujith Nair Smita Kabra

Accenture — Beijing, China — 04/2017-12/2017

Management Consultant Intern - Healthcare Industry

- Collaborated in industry benchmarking, competition analysis for a healthcare and biotechnology valley planning project; Analyzed biomedical industry chain, collected information for industry research report and prepared presentation deck;

IT Consultant Intern - Tobacco Industry, Foodservice Industry

- Designed data analytic plan, visualized simulated sales using Tableau, and forecasted 3 months sales behavior used *regression models* for a tobacco production management project; Collaborated both technology and business tender documents for a digital sales planning platform.

Qinghai University Affiliated Hospital, Information and Network Center — Xining，China — 07/2014-08/2014

Database Administration Intern

- Involved in daily hospital database maintenance;
- Evaluated current Health Information Systems based on national criteria;
- Assisted manager finished demanded analysis for an integrated platform upgrade.

TEACHING ASSISTANT EXPERIENCE

Health Sciences Informatics, Knowledge Engineering and Decision Support — Baltimore, MD — 03/2020-05/2020

- Taught by Dr. Harold Lehmann;
- Organized course material, live talks, provide evaluation, and feedbacks for students
- Assisted quizzes, assignments, and final projects grading.

TIMES: Clinical Informatics Course for Medical Students — Baltimore, MD — 02/2019-04/2019

Taught by Dr. Ashwini Davison

- Created learning objectives, materials, and testing examples for one session.

SKILLS

Language: English (Fluent speaking and writing), Mandarin (Native)

Presentation, Visualization and Communication: PowerPoint, Axure, Tableau, E-Charts

Data Management, Analytics: Oracle 12c, Access, MySQL, Excel, SQL, R, Python, Stata programming,

Web Development: XML, HTML/CSS **Software Development**: Java, Python