Analytic Methods Used In Real World Data Based Biomedical Research- A Scoping Review

by

Chenyu Li, MS

Ray Alsheikh, MD MPH

Karen Robinson, PhD

Harold P Lehmann, MD PhD

# 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, in particular, research based on electronic health record (EHR) data. Because traditional statistics methods used in RWD analysis, may lead to bias, many Real World Methods (RWM) have been recommended. Our objective was to document RWM have been used in EHR-based RWD analysis over the past 10 years and to see if RWM use has changed over that time period,

**Methods:**

We conducted a sampled methodological review of methods used in EHRs based biomedical research in two cohorts. In Cohort 1, we verified the search strategy for papers based on EHER data, validated the data-extraction strategies for RWM, and generated a base rate for sample size consideration for the next cohort. In Cohort 2, we aimed to sample enough papers from each epoch to establish a trajectory of RWE-method use. The primary outcome was proportion of articles using RWM. Meta-regressions were utilized to examine trends in proportion changes over time.  
**Results:**

Epochs were defined as 2010−2013, 2014−2016, 2017, 2018, and 2019. In Cohort 1, 1130 articles were identified, 300 assessed for inclusion, and 88 reviewed in detail. A base rate of .1 was found, yield a sample size of 35 articles per epoch for Cohort 2. In that 10-year cohort, of 175 papers reviewed in detail, 50 (29.%) used recommended 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.% (0.89%, 22.%), 14.% (2.7%, 26.%) , and 49.% (33.%, 65.%), respectively. For RWM, sensitivity analysis, and missing data, meta-regression slopes were statistically 0.  
**Conclusions:** The proportion of the EHR-based studies using RWM or reporting their handling missing data is below 30%; of those reporting using sensitivity analysis, 50%, and the proportions have not changed in the last 10 years. Although regulators, guidelines, books, and academic meetings have attempted to change behavior during the study period, proper analytic methods remain under-used in published studies of EHR-based data.

**Keywords:**

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

# 

# 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](#_ENREF_1)] . There is a growing interest in using RWE in biomedical research by stakeholders, including policymakers, biomedical researchers, clinicians, and medical product developers. [[2-8](#_ENREF_2)] Despite potential broad utility, RWE, compared with Random Controlled Trial (RCT) —still the gold-standard of clinical research[[9](#_ENREF_9)]—is found wanting. In the current “Big Data Era”, investigators are eager to apply Artificial Intelligence, Machine Learning methods in healthcare [ref]. Using an improper or limited method to create RWE from RWD may result in inefficient use of research funds; applying the resulting RWE in real-world practice may result in the improper treatment of patients.

The attention of RWD thus increasingly focuses on data in electronic health records (EHRs). [[10](#_ENREF_10)] 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.”* [[11](#_ENREF_11)] 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[[12](#_ENREF_12)], Use of Electronic Health Record Data in Clinical Investigations-Guidance for Industry[[13](#_ENREF_13)]. 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](#_ENREF_1) [10](#_ENREF_10) [12-14](#_ENREF_12)]

The purpose of this review was to assess the degree to which resarch based on EHR data follow these disseminated guidelines and to assess whether the proportion of such research following these guidelines has improved over the past 10 years.

In creating this report, 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. [[15](#_ENREF_15)] Our Objectives was to document what analytic details investigators performed from RWD to generate RWE, and our Focus was biomedical research papers that used Electronic Health Records as the main data source.

# Methods

We conducted a random-sampled methodological review, for which, we needed (1) to define the articles that would be eligible for review (i.e., operationalize the intention of “papers that used Electronic Health Records as the main data source”), (2) to operationalize the list of RWE methods; and (3) to establish proportions of use of the these methods for time slices over the past 10 years. A search conducted in Cochrane Library and PROSPERO on July 18th, 2021, showed no similar systematic or scoping reviews were registered.

## 1. Eligibility

The intention of our eligibility requirements was for original quantitative research articles publishd 2010−2019 that analyzed data collected from real-world practice to answer biomedical questions written in English. To stabilize our operationalization of this intention and because we discovered a limiting restriction in the NLM-supplied search strategy[[16](#_ENREF_16)] that reduced our yield, we piloted the entire review on one cohort of papers (“Cohort 1”) and executed it on a second (“Cohort 2”).

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. That is, where needed, decisions were aimed at biasing in favor concluding that studies used RWM, so our results would be upper bounds

.

(<Table 1 Search Strategy Inclusion Criteria about here>

We initially created 4 inclusion criteria (see Table 1, numbered 1 to 4) geared towards high sensitively, upon which the search strategy was based.In the course of the reading process, we determined further specifications: No.3b−3d, and No.5-7.

### Search Strategy- cohort 1

We searched peer-reviewed articles in PubMed (MEDLINE). The search term “Electronic Health Records” was implemented through an NLM-supplied search strategy. “Biomedical Research” was operationalized as the PubMed Publication Type, “Study Characteristics.” We excluded review articles. 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.[[18](#_ENREF_18)] We limited the publication date to 2010/01/01-2019/12/31. The detailed search strategy sees Appendix 1.1

### Search Strategy - Cohort 2

A review of the search strategy after completion of Cohort 1 found that the NLM query limited to 8 informatics journals, which is too limiting with regards to our Focus, while the terms for “Biomedical Quantitative Study” were too broad [[19](#_ENREF_19)]. Clinical filters, publication time restriction, and review paper exclusion terms used the same strategy as in the cohort. Cohort 2 search was conducted on 2020/11/09. The detailed search strategy is Appendix 1.2

### Exclusion criteria

We grew the exclusion list in the course of the study. Because of our Focus on routinely-collected Electronic Health Record, we excluded research reports based on claims, genomic, manually-collected Registry, or clinical-trial (RCT) data. Because of our focus on biomedical research on patients, articles that focused on physician behavior, information-system evaluation, health-services evaluation, and new Informtion Technology in healthcare were excluded as well. Finally, research based on unstructured and semi-structured data which needed natural language processing or text mining were excluded. The list was finalized in Cohort 1.

## 2. RWE Methods

From regulatory guidelines, books, and RWD meeting recommendations, we assembled a list of RWM. [[10](#_ENREF_10) [12-14](#_ENREF_12) [20-24](#_ENREF_20)] We developed 3 lists: pre-analytic (missingness), analytic methods, and post-analytic (sensitivity analysis) concerns or methods. These lists, too, were developed in the course of Cohort 1: In the course of reading articles, we encountered novel terms. Where judgment was needed, we biased in favor of ascribing RWM to an article, so, again, our assessments would be an upper bound on the proportion in an epoch using such methods. Thus, machine learning methods combined with causal inference also were considered RWM. [[20](#_ENREF_20)]

## 3. Proportions

We divided the 10 year period, 2010−2019 into 5 epochs: 2010−2013, 2014−2016, 2017, 2018, and 2019, presuming that the proportions would rise, because of guidances that came out during the decade [[10](#_ENREF_10) [12-14](#_ENREF_12) [20](#_ENREF_20) [23](#_ENREF_23)], specifically in the years 2013, 2018, and 2019, and we were hoping to identify impact of any such guidance. Our goal was to distinguish proportions between 2 years. The apparent proportion of key outcomes from Cohort 1 was about 0.1, so we used 0.1 as the estimated proportion, set the desired precision of estimation at 0.1 and confidence level of 0.95 to calculate the sample size for Cohort 2. Using the standard sample-size formula for proportions[[25](#_ENREF_25)], a sample size was calculated for each epoch of 35 per epoch that passed inclusion and exclusion criteria. We sampled articles randomly within each epoch until we reached 35 included articles.

## Analysis and Synthesis process

Citations yielded in the search strategy were downloaded, titles and abstracts were reviewed, and, if not excluded, the full text was reviewed. Characteristics of the study design type, Country/District, mentioned missing data, etc. were documented after reading full-text. The Country/District was defined as the country in which the main population in the database. We documented the methods used in included papers, noting in particular any RWM. The missing data analytics process performed 1. Deletion methods with examining the sensitivity of results to the MCAR and MAR assumptions; [[21](#_ENREF_21)] 2.Single impulation methods; 3. Model based methods , were tagged as handled missing data. [[21](#_ENREF_21) [26](#_ENREF_26)]

All articles were read by a single reader (CL). A second reader RA reviewed a random sample of the articles to check for errors in data documentdation or interpretation. Cohen’s kappa score was calculated to assess agreement[[27](#_ENREF_27)]. Differences of opinion were discussed between the 2 readers and, if necessary, with the senior author (HL). Any questions of the single reader were also discussed among the three. Attention was paid to separate sensitivity analysis, which method was suggested be used whether RWE or more traditional methods are used [[10](#_ENREF_10) [21](#_ENREF_21) [23](#_ENREF_23)] and missing data, which again is a concern in either framework[[10](#_ENREF_10) [13](#_ENREF_13) [20](#_ENREF_20) [21](#_ENREF_21) [23](#_ENREF_23)].

The proportions of papers within each epoch using missing-data methods, RWM, or sensitivity analysis were calculated, along with the confidence intervals, and graphed over time. To assess the impact of any article contexts on rates, we conducted a mixed-effects meta-regression using restricted maximum-likelihood (ReML) using PyMARE, in python,. [[28](#_ENREF_28)] Three meta-regression were done with time as the independent variable. Since the epoch lengths were unequal, we used the midpoint of each epoch in the analysis. To make the intercept meaningful, we subtracted the epoch mid-point (2015) from the epoch values.

# Results

## Study Selection Flow

The study selection flow is summarized using a PRISMA 2009 flow diagram (Figure 1)[[29](#_ENREF_29)].

<Figure 1 about here>

We conducted the literature search on March 23rd, 2020 (Cohort 1) and November 9th ,2020 (Cohort 2). The number of papers retrieved increased, from 1130 in Cohort 1 to 5885 in Cohort 2, confirming our need to proceed with Cohort 2.

To reach the target goal in Cohort 2 of 35 papers per epoch, 392 papers were randomly sampled and reviewed. Reasons for, and numbers of, exclusions are given in Table 2

<Table 2 about here>

The characteristics of included papers are given in Table 3. Of all the included research, 97 (55.% ) were conducted in the United States, 20 (11.%) in the United Kingdom, and 7 (4.0%) both in Korea and China. Of the 175 included research articles, 120 (68.6%) were designed as a retrospective cohort study, 14 (8.0%) were designed as a retrospective cross-sectional study, 14 (8%) were retrospective chart review, and 12 (6.9%) were prospective cohort study . While wide variety of statistical tools were reported, 28(16.0%) reports did not provide any such details.

In Cohort 1, there were 30 papers assessed for inclusion/exclusion with 100% agreement. There were 19 papers assessed for RWM with 100% agreement.

## Proportion Analysis

<Table 6 about here>

<Figure 4 about here>

The raw proportions of each, missing data, RWM, and sensitivity analysis varied between .10 and .4. While the proportions seem to rise by 2017, they return to low values in 2018 and 2019.

## Meta-regressions

The slopes of the regressions over time are numerically close to zero, and are statistically non-significant (see Table 5).

<Table 5 about here>

# Discussion

We took a novel random-sampling approach for this scoping review to assess the proportion of articles that reported methods recommended for RWE. Our results suggest that the proportion of studies using appropriate methods has not changed substantively, whether looking at statistical methods (between 20 and 40%), use of sensitivity analysis (between 9 and 29%), handling missing data (between 14 and 40%) or mentioning missing data (between 37 and 60%)—despite the multiple calls for just such methods[[10](#_ENREF_10) [12-14](#_ENREF_12) [20](#_ENREF_20) [23](#_ENREF_23)]. While there seems to have been a peak in 2017, the 2018 and 2019 rates are no better than 2010−2013. Recalling that we biased our method to be generous, we conclude that the use of RWM in published papers is less than 40%

<Something about how few papers were eligible in the first place>

One of the major challenges in the analysis of EHRs is the missing-data problem [[22](#_ENREF_22) [31](#_ENREF_31)]. Eighty-one (46.2%) included papers mentioned the missing data issue in the data cleaning or limitation session. However, only 39 (22.%) of included papers reported how they handled the missing data or even mentioned missing data at all. A new standard for reporting RWE—STarT-RWE—makes describing how missing data will be handled throughout the analysis and subgroup analyses a requirement. [[32](#_ENREF_32)]If the missingness is Missing at Completely at Random (MCAR) or Missing at Random (MAR), the probability of missing recod is independent of observed data or outcome measurements, dropping the whole record with missing elements would not influence the estimator. However, 19 papers (48.7% among studies that handled missing data ) 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.

The proportion estimation of papers that used RWM in 2018 is 23%, an upper bound of 36%, studies used RWM is disappointingly low. The list of Real-world Methods we used included methods that could analyze the causal effects of observed data, and machine learning methods with proper causal inference.[[20](#_ENREF_20)] 72 (41.1%) papers used traditional regressions and 26 (14.9%) used survival analysis without adjusting baseline confounding and time-varying confounders, as a result, as such analysis have biased outcomes and are limited to inform decision-making[Collaboration, 2020

#1227]. To better interpret and analyze 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[[21](#_ENREF_21) [23](#_ENREF_23)]. In our results, the proportion estimations of studies conducted sensitivity analysis are 0.28(0.14, 0.43), 0.08(0, 0.18), 0.11(0.08, 0.22) in 2017,2018 and 2019 , respectively. STROBE guideline which first published in 2007 suggested that observational studies should describe statistical methods used to control confounding, to examine subgroups and interactions, and to address missing data, also report any sensitivity analyses applied in a study. [[33](#_ENREF_33)] STaRT-RWE suggested in RWE report a separate table should be included for sensitivity analyses, in which investigators may explain which parameters are being modified, why they are being varied, and what they anticipate to learn from this sensitivity study in comparison to the main analysis.[[32](#_ENREF_32)] 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 used at high rates in the published studies, despite guidances over the past 10 years to do so. Why might this low rate be the case? 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 easily reused.

Due to RWD's complexity, it is not appropriate to use traditional data processing methods with large datasets. In order to ensure internal and external validity in EHR-based research, researchers must determine whether the data are accurately extracted, adequately adjusted, correctly analyzed and cogently presented.[[21](#_ENREF_21)] To understand and analyze the RWD in a proper method requires that investigators collaborate in a multidisciplinary team that comprises clinicians, informaticians, epidemiologists, and biostatisticians (data scientists).

The proportion of published studies using RWE methods has not risen over the past 10 years, despite multiple guidances. Alternative approaches are needed.OHDSI (Observational Health Data Sciences and Informatics) has developed tools to conduct real-world evidence generation.[[34](#_ENREF_34)] 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 is comprehensive 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. Wang and colleagues have developed new methodology checklist has been created (STaRT-RWE)[[32](#_ENREF_32)]. We suggest a third alternative is *analytic decision support,* where the analyst receives suggestions from the software in the process of developing an analysis. Using the AHRQ principle of getting the right information to the right user at the right time[[35](#_ENREF_35)], we suggest that such decision support could do for RWE what EHR-based decision support has done for guideline adherence.[ref]

### Limitations

We did not search for all eligible articles. However, our goal was to assess the proportion of such articles using appropriate methods; random sampling is the basis of much biomedical research. The latest 2020 papers were not included in this research. The COVID-19 pandemic may have changed the landscape of analysis; however, while there is evidence that research reported that during COVID-19 has changed, the overall scientific research quality decreased in 2020. [[36-38](#_ENREF_36)] We eliminated 2020 research to reduce the potential bias this real-world situation could bring.

### Conclusion

The proportion of published research papers based on EHR data using real-world evidence methods has not increased significantly over the past 10 years. New approaches are needed to promote analysts’ use of these methods.

.

# Acknowledgements

We gratefully acknowledge the help of Claire Twose, Associate Director for Research Services at Johns Hopkins University, with defining the search strategy and reviewing search implementation.

# 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. Framework for FDA’s Real-World Evidence Program. 2018

11. Congress t. 21st Century Cures Act 2016

12. FDA. Best Practices for Conducting and Reporting Pharmacoepidemiologic Safety Studies Using Electronic Healthcare Data. 2013

13. FDA. Use of Electronic Health Record Data in Clinical Investigations

Guidance for Industry. 2018

14. FDA. Submitting Documents

Using Real-World Data

and Real-World Evidence

to FDA for Drugs and

Biologics

Guidance for Industry. 2019

15. 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]|.

16. MEDLINE / PubMed Search Strategy & Electronic Health Record Information Resources. Secondary MEDLINE / PubMed Search Strategy & Electronic Health Record Information Resources. <https://www.nlm.nih.gov/services/queries/ehr_details.html>.

17. PubMed. Publication Characteristics (Publication Types) with Scope Notes. 2020

18. PubMed. PubMed Clinical Queries 2020

19. NLM. Epidemiologic Study Characteristics. Secondary Epidemiologic Study Characteristics 2020. <https://www.ncbi.nlm.nih.gov/mesh/68016020>.

20. FDA. Developing Real-World Data and Evidence to Support Regulatory Decision-Making, 2019.

21. MIT. Secondary Analysis of Electronic Health Records. 2016

22. 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]|.

23. 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]|.

24. 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]|.

25. Pourhoseingholi MA, Vahedi M, Rahimzadeh M. Sample size calculation in medical studies. Gastroenterol Hepatol Bed Bench 2013;**6**(1):14-17

26. Holmes JH. Knowledge Discovery in Biomedical Data: Theory and Methods. Methods in Biomedical Informatics, 2014:179-240.

27. McHugh ML. Interrater reliability: the kappa statistic. Biochemia medica 2012;**22**(3):276-82

28. PyMARE: Python Meta-Analysis & Regression Engine [program], 2020.

29. 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]|.

30. McHugh ML. Interrater reliability: the kappa statistic. Biochem Med (Zagreb) 2012;**22**(3):276-82

31. 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]|.

32. Wang SV, Pinheiro S, Hua W, et al. STaRT-RWE: structured template for planning and reporting on the implementation of real world evidence studies. BMJ 2021;**372**:m4856 doi: 10.1136/bmj.m4856[published Online First: Epub Date]|.

33. von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) Statement: Guidelines for Reporting Observational Studies. Annals of Internal Medicine 2007;**147**(8):573-77 doi: 10.7326/0003-4819-147-8-200710160-00010[published Online First: Epub Date]|.

34. Ryan. P DJ. The Book of OHDSI. The book of OHDSI -Observational Health Data Sciences and Informatics, 2020.

35. AHRQ. Overview of CDS Five Rights Secondary Overview of CDS Five Rights <https://digital.ahrq.gov/ahrq-funded-projects/current-health-it-priorities/clinical-decision-support-cds/chapter-1-approaching-clinical-decision/section-2-overview-cds-five-rights>.

36. Alexander PE, Debono VB, Mammen MJ, et al. COVID-19 coronavirus research has overall low methodological quality thus far: case in point for chloroquine/hydroxychloroquine. Journal of Clinical Epidemiology 2020;**123**:120-26 doi: 10.1016/j.jclinepi.2020.04.016[published Online First: Epub Date]|.

37. Celi LA, Charpignon M-L, Ebner DK, et al. Gender Balance and Readability of COVID-19 Scientific Publishing: A Quantitative Analysis of 90,000 Preprint Manuscripts: Cold Spring Harbor Laboratory, 2021.

38. Dobler CC. Poor quality research and clinical practice during COVID-19. Breathe 2020;**16**(2):200112 doi: 10.1183/20734735.0112-2020[published Online First: Epub Date]|.

39. Madigan D, Stang PE, Berlin JA, et al. A Systematic Statistical Approach to Evaluating Evidence from Observational Studies. Annual Review of Statistics and Its Application 2014;**1**(1):11-39 doi: 10.1146/annurev-statistics-022513-115645[published Online First: Epub Date]|.

40. Hebert C, Shivade C, Foraker R, et al. Diagnosis-specific readmission risk prediction using electronic health data: a retrospective cohort study. BMC Med. Inform. Decis. Mak. 2014;**14**:65 doi: 10.1186/1472-6947-14-65[published Online First: Epub Date]|.

# Tables and Figures

Table . Search Strategy Inclusion criteria\*

|  |  |
| --- | --- |
| Criterion Number | Criterion |
| 1 | Original Data |
| 2 | Quantitative Study |
| 3 | 1. Allow other source of data combined with the EHRs 2. Using the EHRs as the main source of data for analysis 3. Allow collections of EHR data 4. Either variables or outcomes should come from EHR |
| 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 |

Developed during Cohort 1

Table . Reasons for excluding articles

|  |  |  |
| --- | --- | --- |
|  | Exclusion reason | Number Excluded |
| 1 | Physician behavior, system evaluation, health services research [I.e., not biomedical] | 59 |
| 2 | EMR data only used to identify the cohort | 35 |
| 3 | Technology question(Database build, data collection, datatransmission, IT infrastruction ) | 27 |
| 4 | Registry (data where a human being has abstracted the data [adds data quality; avoid curated data]) | 20 |
| 5 | Questionnaire/survey only | 19 |
| 6 | Methodology papers | 14 |
| 7 | Predictive models | 8 |
| 8 | Patient generated health data only | 7 |
| 9 | RCT data (data where a human being has abstracted the data [adds data quality; avoid curated data]) | 6 |
| 10 | Not English | 6 |
| 11 | Text mining / NLP | 4 |
| 12 | Claim data only | 4 |
| 13 | Genomic data | 3 |
| 14 | Review papers | 2 |
| 15 | Qualitative data only | 1 |
|  | **Total Count** | **215** |

Table Included Paper Characteristics

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Grouped by Epoch | | | | | | | | |
|  |  | **Missing** | **Overall** | **2010-2013** | **2014-2016** | **2017** | **2018** | **2019** |
| Total papers, n |  |  | **175** | **35** | **35** | **35** | **35** | **35** |
| Study Design Type,  n (%) | **retrospective chart review** | 0 | 1 (0.6) |  |  |  |  | 1 (2.9) |
| **cost-benefit analysis** |  | 1 (0.6) |  |  |  |  | 1 (2.9) |
| **prospective cohort study** |  | 12(6.9) | 4 (11.4) |  | 4 (11.6) | 1 (2.9) | 3 (8.6) |
| **prospective controlled study** |  | 1 (0.6) | 1 (2.9) |  |  |  |  |
| **retrospective case–control study** |  | 8 (4.6) |  | 3 (8.6) | 2 (5.7) |  | 3 (8.6) |
| **retrospective chart review** |  | 14 (8.0) | 2 (5.7) | 4 (11.4) | 2 (5.7) | 4 (11.4) | 2 (5.7) |
| **retrospective cohort study** |  | 120 (68.6) | 24 (68.6) | 25 (71.4) | 23 (65.7) | 26 (74.3) | 22 (62.9) |
| **retrospective cross-sectional study** |  | 14 (8.0) | 4 (11.4) |  | 4 (11.4) | 4 (11.4) | 2 (5.7) |
| **retrospective database study** |  | 1 (0.6) |  |  |  |  | 1 (2.9) |
| **retrospective review** |  | 3 (1.7) |  | 3 (8.6) |  |  |  |
| Country/district ,  n (%) | **Australia** | 0 | 3 (1.7) | 1 (2.9) |  |  |  | 2 (5.7) |
| **Brazil** |  | 1 (0.6) |  |  |  | 1 (2.9) |  |
| **Canada** |  | 3 (1.7) | 3 (8.6) |  |  |  |  |
| **China** |  | 7 (4.0) |  | 2 (5.7) |  | 3 (8.6) | 2 (5.7) |
| **Croatia** |  | 1 (0.6) |  |  |  |  | 1 (2.9) |
| **Denmark** |  | 1 (0.6) |  |  |  |  | 1 (2.9) |
| **France** |  | 5 (2.9) |  |  | 2 (5.7) | 2 (5.7) | 1 (2.9) |
| **Israel** |  | 6 (3.4) |  | 2 (5.7) | 1 (2.9) | 1 (2.9) | 2 (5.7) |
| **Italy** |  | 1 (0.6) |  | 1 (2.9) |  |  |  |
| **Japan** |  | 3 (1.7) |  |  |  |  | 3 (8.6) |
| **Japan** |  | 1 (0.6) |  |  |  | 1 (2.9) |  |
| **Korea** |  | 7 (4.0) | 2 (5.7) | 3 (8.6) | 2 (5.7) |  |  |
| **Malawi** |  | 1 (0.6) |  |  | 1 (2.9) |  |  |
| **Mexico** |  | 1 (0.6) |  | 1 (2.9) |  |  |  |
| **Netherlands** |  | 1 (0.6) |  | 1 (2.9) |  |  |  |
| **Oman** |  | 2 (1.1) |  | 1 (2.9) | 1 (2.9) |  |  |
| **Portugal** |  | 1 (0.6) |  | 1 (2.9) |  |  |  |
| **Singapore** |  | 4 (2.3) |  |  |  | 3 (8.6) | 1 (2.9) |
| **Spain** |  | 4 (2.3) | 1 (2.9) | 1 (2.9) |  | 1 (2.9) | 1 (2.9) |
| **Sweden** |  | 2 (1.1) | 1 (2.9) |  |  |  | 1 (2.9) |
| **Switzerland** |  | 1 (0.6) |  |  | 1 (2.9) |  |  |
| **Taiwan** |  | 1 (0.6) |  |  | 1 (2.9) |  |  |
| **Turkey** |  | 1 (0.6) |  |  |  |  | 1 (2.9) |
| **UK** |  | 20 (11.4) | 2 (5.7) | 3 (8.6) | 8 (22.9) | 4 (11.4) | 3 (8.6) |
| **USA** |  | 97 (55.4) | 25 (71.4) | 19 (54.3) | 18 (51.4) | 19 (54.3) | 16 (45.7) |
| Mention Missing Data,  n (%) | **No** | 0 | 94 (53.7) | 19 (54.3) | 23 (65.7) | 14 (40.0) | 20 (57.1) | 18 (51.4) |
| **Yes, data cleaning** |  | 34 (19.4) | 6 (17.1) | 6 (17.1) | 11 (31.4) | 6 (17.1) | 5 (14.3) |
| **Yes, data description** |  | 25 (14.3) | 5 (14.3) | 3 (8.6) | 6 (17.1) | 5 (14.3) | 6 (17.1) |
| **Yes, limitation** |  | 20 (11.4) | 5 (14.3) | 3 (8.6) | 4 (11.4) | 3 (8.6) | 5 (14.3) |
| **Yes, no missing** |  | 2 (1.1) |  |  |  | 1 (2.9) | 1 (2.9) |
| Handled Missing Data  , n (%) | **No** | 0 | 136 (77.7) | 27 (77.1) | 28 (80.0) | 22 (62.9) | 29 (82.9) | 30 (85.7) |
| **Yes, Imputation** |  | 17 (9.7) | 3 (8.6) | 2 (5.7) | 5 (14.3) | 4 (11.4) | 3 (8.6) |
| **Yes, excluded** |  | 19 (10.9) | 4 (11.4) | 5 (14.3) | 6 (17.1) | 2 (5.7) | 2 (5.7) |
| **Yes, sensitivity analysis** |  | 3 (1.7) | 1 (2.9) |  | 2 (5.7) |  |  |
| Followed Check List,  n (%) | **No** | 0 | 171 (97.7) | 34 (97.1) | 34 (97.1) | 34 (97.1) | 35 (100.0) | 34 (97.1) |
| **STROBE** |  | 4 (2.3) | 1 (2.9) | 1 (2.9) | 1 (2.9) |  | 1 (2.9) |
| Analytic Tools Used ,  n (%) | **CART Salford Predictive Miner** | 0 | 1 (0.6) | 1 (2.9) |  |  |  |  |
| **Didn't mention** |  | 28 (16.0) | 6 (17.1) | 5 (14.3) | 8 (22.9) | 3 (8.6) | 6 (17.1) |
| **EZR** |  | 1 (0.6) |  |  |  |  | 1 (2.9) |
| **Excel** |  | 1 (0.6) |  |  |  |  | 1 (2.9) |
| **GraphPad Prism** |  | 1 (0.6) |  |  |  |  | 1 (2.9) |
| **JMP** |  | 1 (0.6) | 1 (2.9) |  |  |  |  |
| **Mplus** |  | 1 (0.6) | 1 (2.9) |  |  |  |  |
| **NCSS** |  | 1 (0.6) |  |  |  |  | 1 (2.9) |
| **MedCalc** |  | 1 (0.6) |  | 1 (2.9) |  |  |  |
| **Epidata** |  | 1 (0.6) |  |  |  |  | 1 (2.9) |
| **R** |  | 20 (11.4) |  | 2 (5.7) | 5 (14.3) | 9 (25.7) | 4 (11.4) |
| **SAS** |  | 41 (23.4) | 8 (22.9) | 8 (22.9) | 10 (28.6) | 10 (28.6) | 5 (14.3) |
| **SigmaPlot** |  | 1 (0.6) |  |  |  |  | 1 (2.9) |
| **SPSS ( & PASW Statistics)** |  | 50 (28.6) | 11 (31.4) | 13 (37.1) | 10 (28.6) | 10 (28.6) | 6 (17.1) |
| **Stata** |  | 37 (21.1) | 7 (20.0) | 10 (28.6) | 3 (8.6) | 8 (22.9) | 9 (25.7) |
| **Statistica** |  | 1 (0.6) | 1 (2.9) |  |  |  |  |
| **Statview** |  | 1 (0.6) | 1 (2.9) |  |  |  |  |

Table . Proportions (and confidence intervals) of key methodologies, by epoch

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | 2010-2013 | 2014-2016 | 2017 | 2018 | 2019 |
| Real-World\_Method | 0. 23  (0.01, 0.04) | 0.37  (0.21, 0.53) | 0.40  (0.24, 0.56) | 0.23  (0.09, 0.38) | 0.20  (0.07, 0.33) |
| Sensitivity\_Analysis | 0.11  (0.09, 0.22) | 0.29  (0.14, 0.43) | 0.29  (0.14, 0.44) | 0.09  (0, 0.18) | 0.11  (0.09, 0.22) |
| Handled\_Missing\_Data | 0.23  (0.09, 0.37) | 0.20  (0.07, 0.33) | 0.40  (0.24, 0.56) | 0.23  (0.09, 0.37) | 0.14  (0.03, 0.25) |
| Mentioned\_Missing\_Data | 0.46  (0.29, 0.62) | 0.37  (0.21. 0.53) | 0.60  (0.44, 0.76) | 0.43  (0.26, 0.59) | 0.49  (0.32, 0.65) |

Table Meta-regression for three methods

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| name | estimate | se | z-score | p-value | ci\_0.025 | ci\_0.975 |
| intercept | 0.32 | 0.36 | 0.89 | 0.37 | -0.39 | 1.03 |
| Real-World Methods | 0.00 | 0.02 | -0.14 | 0.89 | -0.05 | 0.04 |
| intercept | 6.42 | 3.29 | 0.19 | 0.85 | -58.83 | 71.68 |
| Sensitivity Analysis | 0.00 | 0.02 | -0.19 | 0.85 | -0.04 | 0.03 |
| intercept | 8.93 | 41.75 | 0.21 | 0.83 | -72.89 | 90.75 |
| Handle Mssing Data | 0.00 | 0.02 | -0.21 | 0.83 | -0.04 | 0.04 |

Footnote

The meta-regression were done with time ( middle point of each epoch) as the independent variable. The hypothesis here is the proportion of use RWE, conduct sensitivity analysis, and handle missing data changed year from year. As the result shown in the table, proportion of used RWM did not increase from year to year, the upper bound is increase 4%. Proportion of conducted sensitivity analysis did not increase, the upper bound is increase 3%,

| 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 |

Cohort 1

Table Proportion estimation and Confidence Interval



Figure PRISMA Flow Diagram

Chart, bar chart

Description automatically generated

Figure . Number of articles retrieved, by year

Chart, bar chart

Description automatically generated

Figure 3 Included Papers by Epochs

Chart, box and whisker chart

Description automatically generated

Figure 4 Proportion of Methods Used in the RWD Resesarch

<Cohort 1 graphs>

1. 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) |

1. Included papers by epoch

A screenshot of a cell phone

Description automatically generated

1. 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) |

1. Proportion of Methods Used

A screenshot of a cell phone

Description automatically generated

# Appendix

Appendix .1 Search Strategy Details Cohort1( on March 23rd, 2020)

|  |  |  |
| --- | --- | --- |
| Intention | Operationalization | 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  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 [[39](#_ENREF_39)] OR  j am med inform assoc [[39](#_ENREF_39)] OR  amia annu symp proc [[39](#_ENREF_39)] OR  health data manag [[39](#_ENREF_39)] OR  int j med inform [[39](#_ENREF_39)] OR  yearb med inform [[39](#_ENREF_39)] OR  telemed j e health [[39](#_ENREF_39)] OR  stud health technol inform [[39](#_ENREF_39)]) | MEDLINE / PubMed Search Strategy & Electronic Health Record Information Resources  https://www.nlm.nih.gov/services/  queries/ehr\_details.html |
| 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 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]) | Clinical Queries using Research Methodology Filters  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 1.2 Search Strategy Details Cohort2 ( on November 9th, 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  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]) | MEDLINE / PubMed Search Strategy & Electronic Health Record Information Resources  https://www.nlm.nih.gov/services/  queries/ehr\_details.html |
| Biomedical Quantitative Study  (#2) | "Epidimeological Study Characteristics"[[40](#_ENREF_40)] AND “data”[All fileds] AND “analy\*”[All Fields] NOT "Review"[Publication Type] NOT “Systematic Review"[Publication Type] | Works about types and formulations of studies used in epidemiological research.  Year introduced: 2018 (1998)  https://www.ncbi.nlm.nih.gov/mesh/68016020 |
| 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 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]) | Clinical Queries using Research Methodology Filters  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 PubMed searched result

[PubMed file -github](https://github.com/ChenyuL/ANALYTIC-METHODS-USED-IN-REAL-WORLD-DATA-BASED-BIOMEDICAL-RESEARCH/blob/master/pubmed-healthinfo-set.nbib)

Appendix EndNote Library

[EndNote Library-github](https://github.com/ChenyuL/ANALYTIC-METHODS-USED-IN-REAL-WORLD-DATA-BASED-BIOMEDICAL-RESEARCH/blob/master/Export_Articles_EndNote.pdf)

Appendix Excel Database

[Excel-Database-github](https://github.com/ChenyuL/ANALYTIC-METHODS-USED-IN-REAL-WORLD-DATA-BASED-BIOMEDICAL-RESEARCH/blob/master/Appendix-Excel_Database.xlsx)

Table 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](https://github.com/ChenyuL/ANALYTIC-METHODS-USED-IN-REAL-WORLD-DATA-BASED-BIOMEDICAL-RESEARCH/blob/master/Analytics_Code.ipynb)

# <from - RWM-appendix >

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| RWM | cohrot 1/2 | NonRWM | Cohort Number |  |
| Confounding | 1 |  | Mixed -effects Regression Model | 1 |
| Non-adherence | 1 |  | Logistic Regression | 1 |
| Immortal Time | 1 |  | Multivariate logistic regression | 1 |
| Causal Inference | 1 |  | Empirical Bayes estimates | 1 |
| Inverse Probability | 1 |  | Kaplan-Meier method | 1 |
| Adjusting | 1 |  | Linear Regression | 1 |
| Bias | 1 |  | Cox proportional hazard models | 1 |
| Sensitivity Analysis | 1 |  | Discriminant function analysis | 1 |
| Trimming | 1 |  | Individual growth curve (IGC) analysis | 1 |
| Propensity Score | 1 |  | Descriptive statistics | 1 |
| Instrumental Variable | 1 |  | Intention-to-Treat Analysis | 1 |
| G-Estimation | 1 |  | retrospective chart review | 1 |
| Marginal Structure Models | 1 |  | Post-hoc anaysis | 1 |
| Doubly Robust Methods | 1 |  | Multivariate linear regression | 1 |
| Targeted Maximum Likelihood Estimation | 1 |  | Hypothesis test | 1 |
| Active Comparator | 1 |  | one-way analysis of variance (ANOVA) | 1 |
| Negative Control | 1 |  | Multiple logistic regression | 1 |
| High-dimentional Proxy Adjustment | 1 |  | Linear mixed effect model | 1 |
| Reverse Causation | 1 |  | LASSO regression | 1 |
| Depletion of Susceptible | 1 |  | Multilevel logistic regression | 1 |
| Pseudo Treatment | 1 |  | Generalized Estimating Equations | 1 |
| (Manski's) Partial Identification | 1 |  | Poisson regression | 1 |
| Empirical Calibration | 1 |  | recursive partitioning (RP) model | 1 |
| Regression Discontinuity | 1 |  | C-Statistics | 1 |
| Missing Cause | 1 |  | Decision Curve Analysis | 1 |
| Perturbation Variable | 1 |  | Random Forest | 1 |
| Difference in Difference | 1 |  | multivariable generalized linear mixed model | 1 |
| Trend in Trend | 1 |  | Chi-Square Automatic Interaction Detector | 1 |
| Bayesian Twin Regression | 1 |  | multiple correspondence analysis | 1 |
| Multiple Imputation | 1 |  | Hierarchic cluster analysis | 1 |
| DAG/ADMG | 1 |  | Difference-in-difference analysis | 1 |
| Identification | 1 |  | group-based multitrajectory analysis | 1 |
| Missing Data | 1 |  | Optimal Classification Trees | 1 |
| latent class growth modeling (LCGM) | 1 |  | AdaBoost | 1 |
|  |  |  | multilevel quantile regression | 1 |
|  |  |  | Statistic Testing | 1 |
|  |  |  | Long Short Term Memory (LSTM) Network | 2 |
|  |  |  | Gamma Regression | 2 |
|  |  |  | Multivariate analysis of variance (MANOVA) | 2 |
|  |  |  | hierarchical generalized linear model | 2 |
|  |  |  | interrupted time series analysis | 2 |
|  |  |  | Growth mixture modelling | 2 |
|  |  |  | multivariable negative binomial regression | 2 |
|  |  |  | Subgroup Analysis | 2 |
|  |  |  | life-table analysis | 2 |