JOHNS HOPKINS
BLOOMBERG SCHOOL
of PUBLIC HEALTH

Overview of Data and Methods Used in Pharmacoepidemiology

Hemalkumar B. Mehta, MS, PhD
Johns Hopkins University

Produced by the Center for Teaching and Learning at the Johns Hopkins Bloomberg School of Public Health.

Contents

- Randomization vs. the real world
- Data sources
- Study designs
- Statistical analysis
- Common statistical software used in pharmacoepidemiology
- Checklist and protocols

Randomization vs. the Real World

Produced by the Center for Teaching and Learning at the Johns Hopkins Bloomberg School of Public Health.

Randomized Trials (Experimental) vs. Real-World Studies (Observational)

Randomized trials (experimental)

- Planned and well-controlled experiments
- Randomization
- Consistent monitoring and follow-up
- Measures efficacy
- Prerequisite for drug approval
- Gold standard or highest in the hierarchy of study designs

Real-world studies (observational)

- What happens in everyday clinical practice
- Lacks randomization
- No strict protocol for treatment and follow-up
- Measures effectiveness
- Can be used to assess use, safety, and effectiveness in the real world

Randomized Controlled Trials Are Not Always Ideal

- Not ideal for studying real-world effectiveness of a drug
- Not ideal for studies of harms of drugs (not usually big enough and with long follow-up)
- Difficult to generalize findings to large groups of people
- Hard to look at subgroups with sufficient power

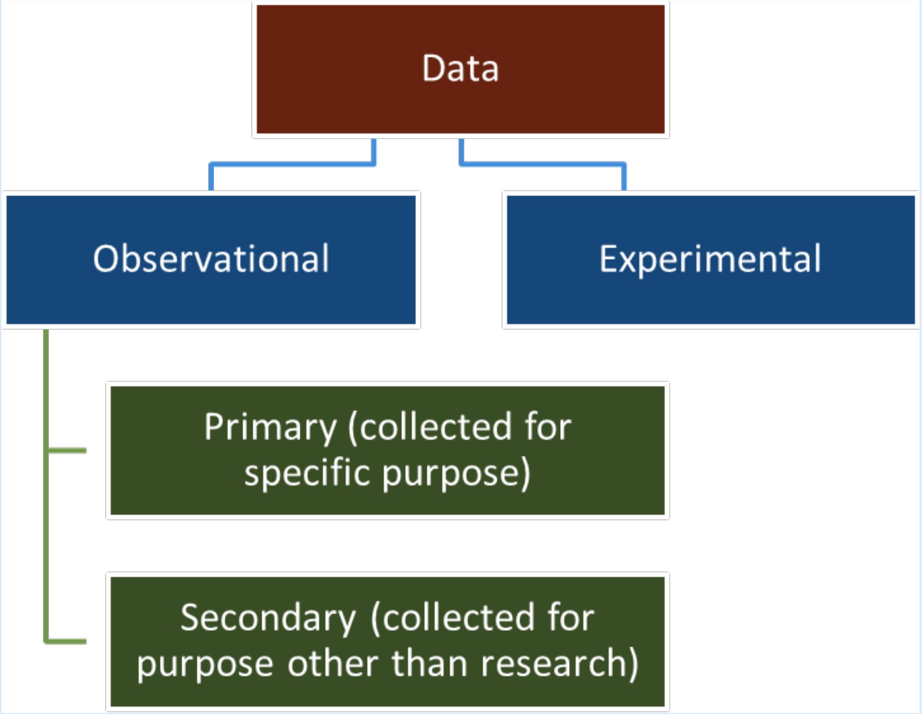
- Limited to interests of stakeholders who fund them
- Can answer only one question at a time, usually
- Long (multiple sites, multiple investigators, rare patients)
- Expensive

Real-World Studies

- Can describe outcomes of patients receiving care in their usual care setting from their usual clinicians
- Can be designed to answer many questions
- Can sometimes be large without much increased expense over a small study, depending on design
- Can more easily report on safety, harms, adverse events, drug–drug interactions, and discontinuation of medication
- Need data and analytic tools to answer important questions from real-world data

Data Sources

Common Types of Data Sources


Data
Observational
Experimental
Primary (collected for specific purpose)
Secondary (collected for purpose other than research)

- Primary data source, used for
  1. Registry data
  2. Survey data
  3. Prospective cohort studies

- Secondary data source (used for majority of real-world studies)
  4. Administrative claims data
  5. Electronic health records data

Image source: Hemalkumar B. Mehta, with the Center for Teaching and Learning, Johns Hopkins Bloomberg School of Public Health.

1. Registry Data

- Organized system that collects clinical and other data in a standardized format for a population defined by a particular disease, condition, or drug exposure

- One example: Surveillance, Epidemiology, and End Result (SEER) registry for cancer
  - Collects cancer incidence and survival data on US population
  - Provides comprehensive cancer statistics in the US
  - Can be linked with other datasets

- Use cases
  - Characterize natural history of disease
  - Biomarker discovery
  - Characterize effectiveness and safety
  - National statistics
  - Clinical trials
    - Selection of study individuals
    - Endpoint selection

Sources:
https://www.fda.gov/media/154449/download
https://www.nih.gov/health-information/nih-clinical-research-trials-you/list-registries
https://seer.cancer.gov/about/overview.html

2. Survey Data

- National surveys to collect information on health, nutrition, cost, etc.
- Several national surveys conducted by:
  - National Center for Health Statistics (NCHS)
  - Center for Disease Control and Prevention (CDC)
  - Agency for Healthcare Research and Quality (AHRQ)
- Mostly freely available

- One example: Medical Expenditure Panel Survey (MEPS)
  - Surveys of families and individuals, their medical providers (doctors, hospitals, pharmacies, etc.), and employers
  - Collects information on
    - Health services use
    - Clinical conditions
    - Prescription drugs
    - Cost

Sources:
https://www.cdc.gov/nchs/surveys/index.html
https://meps.ahrq.gov/mepsweb/

3. Prospective Cohort Study

- Enroll a group of people (cohort) and follow them over time to collect health and other information using questionnaires, biological assays, interview, lab measurements

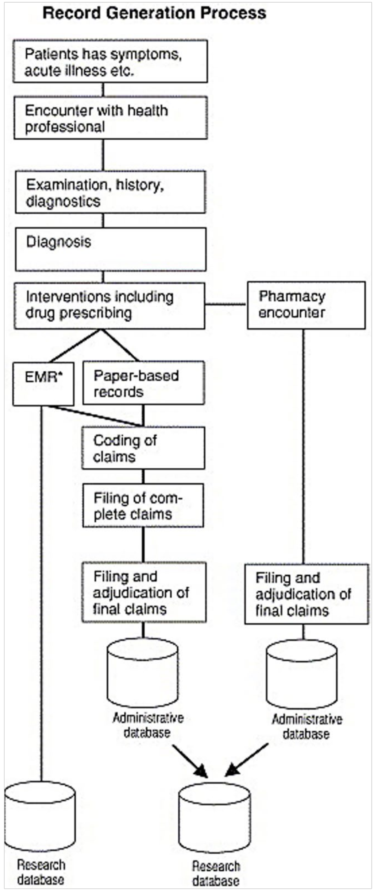
- Several prospective cohort studies:
  - Framingham Study
  - Baltimore Longitudinal Study on Aging
  - Women's Health Initiative Study

- One example: Framingham Study
  - Began in 1948
  - Aimed to unravel the underlying causes of heart disease
  - Significantly improved our understanding of heart disease and informed the current clinical practice
  - Major public health impacts

Sources:
https://iadrp.nia.nih.gov/about/cadro/Population-Studies-Cohorts-and-Studies
https://www.nih.gov/sites/default/files/about-nih/impact/framingham-heart-study.pdf
https://www.framinghamheartstudy.org/

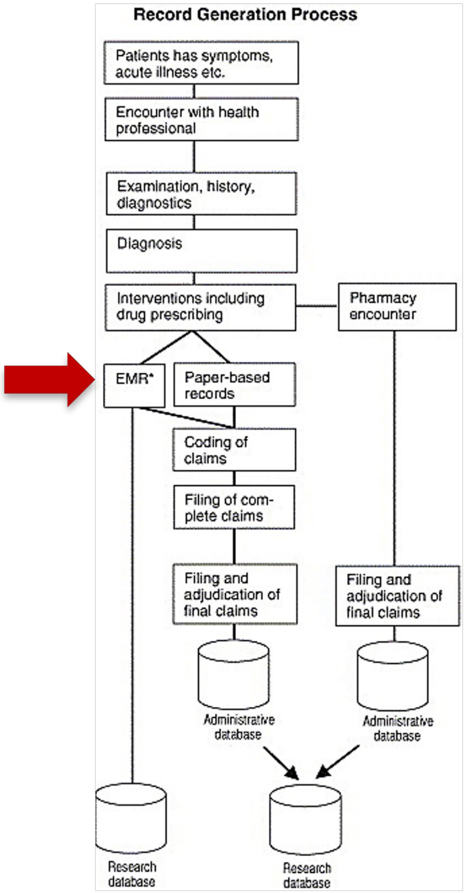
Administrative claims data and electronic medical records data are widely used for pharmacoepidemiologic research

Record Generation Process


Record Generation Process
Patients has symptoms, acute illness etc.
Encounter with health professional
Examination, history, diagnostics
Diagnosis
Interventions including drug prescribing
Pharmacy encounter
EMR*
Paper-based records
Coding of claims
Filing of complete claims
Filing and adjudication of final claims
Filing and adjudication of final claims
Administrative database
Administrative database
Research database
Research database

Source: Schneeweiss S, Avorn J. A review of uses of health care utilization databases for epidemiologic research on therapeutics. *J Clin Epidemiol*. 2005;58(4):323-337. doi:10.1016/j.jclinepi.2004.10.012

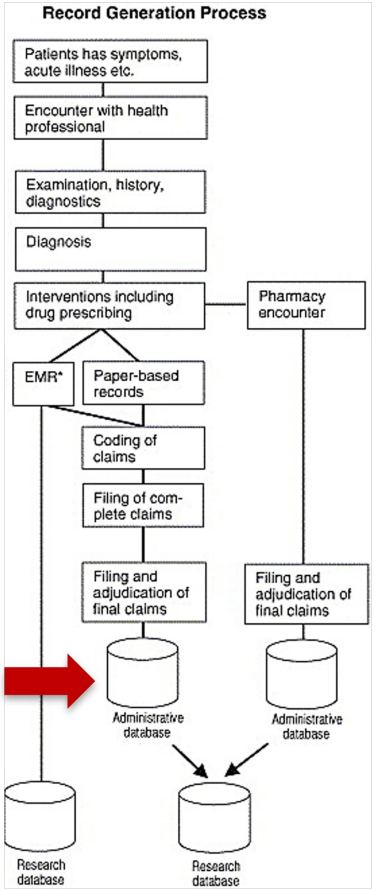
Electronic Medical Records Data


Record Generation Process
Patients has symptoms, acute illness etc.
Encounter with health professional
Examination, history, diagnostics
Diagnosis
Interventions including drug prescribing
Pharmacy encounter
EMR*
Paper-based records
Coding of claims
Filing of complete claims
Filing and adjudication of final claims
Filing and adjudication of final claims
Administrative database
Administrative database
Research database
Research database

- Text describing encounters (history and physical examination findings)
- Text describing procedures
- Text describing behaviors
- Text describing prescriptions written
- From inpatient encounters
- From outpatient encounters
- Laboratory results
- Radiology results (text)
- Problem list
- Active medication list
- Archived medication list
- Provider
- Age, sex, race/ethnicity
- Insurance information

Source: Schneeweiss S, Avorn J. A review of uses of health care utilization databases for epidemiologic research on therapeutics. *J Clin Epidemiol*. 2005;58(4):323-337. doi:10.1016/j.jclinepi.2004.10.012

Administrative Data From Insurance Companies


Record Generation Process
Patients has symptoms, acute illness etc.
Encounter with health professional
Examination, history, diagnostics
Diagnosis
Interventions including drug prescribing
Pharmacy encounter
EMR*
Paper-based records
Coding of claims
Filing of complete claims
Filing and adjudication of final claims
Filing and adjudication of final claims
Administrative database
Administrative database
Research database
Research database

- Unique patient number
- Enrollment dates
- Procedure and diagnostic codes
- Dates of service
- Age, sex
- Reimbursed costs
- Sometimes out-of-pocket costs
- Place of service
- Provider type
- Type of admission
- Discharge destination
- Pharmacy claims (drug dispensed and dose, pill count, day supply, date)

Source: Schneeweiss S, Avorn J. A review of uses of health care utilization databases for epidemiologic research on therapeutics. *J Clin Epidemiol*. 2005;58(4):323-337. doi:10.1016/j.jclinepi.2004.10.012

4. Administrative Claims Data—1

- Generated by the insurance company when they are billed by a service provider

- Separate data files for different health care services
  - Enrollment
  - Inpatient
  - Outpatient
  - Prescription drug
  - Skilled nursing facility

- Examples of claims datasets:
  - Medicare
  - Medicaid
  - BlueCross/Blue Shield
  - UnitedHealthcare

4. Administrative Claims Data—2

Strengths

- Records all interactions with health care provider and captures health care received across the care continuum
- Complete information on diagnoses, procedures, prescription drug dispensing, and payment
- Typically, large samples of patients
- Relatively low cost
- High data quality due to financial transactions

Weaknesses

- Missing information on nonprescription medications and prescriptions paid out of pocket
- Lack of detailed clinical information
- Lack of information on lifestyle variables and health behaviors
- Lag time due to adjudication process

4. Medicare: Administrative Claims Data

- Accessed through a contractor: ResDAC (resdac.org)
  - Research Identifiable Files—contains patient identifiable information
  - Limited Data Sets—limited patient identifiable information
  - Public Use Files—aggregate level with no patient identifiable information

- The Standard Analytical Files (SAFs) contain information collected by Medicare to pay for health care services provided to a Medicare beneficiary

- SAFs are available for:
  - Institutional services
    - Inpatient
    - Outpatient
    - Skilled nursing facility
    - Hospice
    - Home health agency
  - Noninstitutional services
    - Physician
    - Durable medical equipment

Source: https://resdac.org/articles/differences-between-rif-lds-and-puf-data-files

5. Electronic Health Records Data—1

- Generated when individual visits health care provider or facility to seek medical care

- Medical records to track information about patients
  - Individual visits primary care physician or specialist for medical conditions
  - Individual visits emergency room or hospital for medical conditions
  - Not meant for billing or insurance purpose

- Examples of electronic health records datasets:
  - Clinical Practice Research Datalink
  - The Health Improvement Network
  - Geisinger
  - Veterans Affairs
  - Electronic health records from Johns Hopkins

5. Electronic Health Records Data—2

Strengths

- Contains rich data on clinical variables such as laboratory results, physiological measurements, and vitals
- Might provide rationale for treatment decision, depending on quality of free text
- Might capture over-the-counter medication use
- Detailed information provided from hospital stays

Weaknesses

- Variability in quality of data
- Incomplete data is a common challenge
- No universal standard for types of data in electronic health record (EHR), resulting in heterogeneity among EHR systems
- Unable to determine whether absence of data is due to care received outside of health care or truly no care was received

5. CPRD: Electronic Health Records Data

- Accessed through Clinical Practice Research Datalink (CPRD; cprd.com)
  - Longitudinal primary care records
  - 2,000 primary care practices and includes 60 million patients
    - 18 million active patients with at least 20 years of follow-up for 25% of the patients

- The CPRD contains anonymized EHR data from primary care practices:
  - Demographic characteristics
  - Diagnoses and symptoms
  - Drug exposures
  - Vaccination history
  - Laboratory tests
  - Referrals to hospital and specialist care

- Can be linked with other datasets

Source: https://www.cprd.com/data

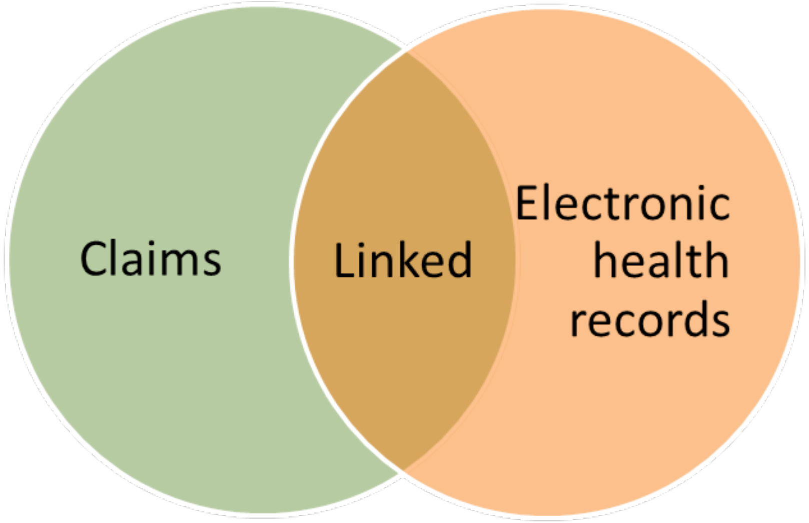
Claims vs. Electronic Health Records Data—1

Information	Electronic Health Records	Administrative Claims
Age/sex	Y	Y
Race/ethnicity	Possibly	Possibly
Socioeconomic data	Possibly	N (maybe ZIP code level)
Insurance information	Y	Y
Prescriptions ordered	Y	N
Pharmacy data (drugs dispensed)	Y or N	Y
Clinical data: vital signs or point-of-care testing results	Y	N
Clinical data: inpatient	Y	N
Clinical data: outpatient	Y	N

Claims vs. Electronic Health Records Data—2

Information	Electronic Health Records	Administrative Claims
Spontaneously reported adverse events	Y (if sought medical care)	Y (if sought medical care)
Diagnoses or procedures coded for payment	N	Y
Behavioral risk factors and diet	Y	N
Indicators of procedures having being done (laboratory, radiological, procedures)	Y	Y
Results from procedures (echocardiography, radiology)	Y	N
Laboratory results	Y	N
Problem list or summary	Y	N
Prescriptions ordered	Y	N
Pharmacy data (drugs dispensed)	N	Y

Electronic Health Record + Claims Data



- Both types of data can complement each other
  - Improve validity
  - Broaden the scope of research
- If all data can be linked (complete overlap), then analysis can be done in linked dataset
- If some data can be linked (partial overlap), then analysis can be done in claims data and enhance analyses in the subset with EHR

Image source: Hemalkumar B. Mehta.
Source: Schneeweiss S, Desai RJ, Ball R. A future of data-rich pharmacoepidemiology studies: transitioning to large-scale linked electronic health record + claims data. *Am J Epidemiol*. 2025;194(2):315-321. doi:10.1093/aje/kwae226

Example of Pharmacoepi Studies

Administrative claims data

- Use of oral anticoagulants among individuals with cancer and atrial fibrillation in the United States, 2010–2016
  - Surveillance, Epidemiology, and End Results (SEER)–Medicare data
  - Characterized drug use among individuals with cancer and atrial fibrillation

Electronic health records

- Glucagon-like peptide-1 receptor agonists (GLP-1) and risk of suicidality among patients with type 2 diabetes
  - UK Clinical Practice Research Datalink linked to the Hospital Episodes Statistics
  - Assessed the use of GLP-1 and risk of suicide

Sources:

Ardeshirrouhanifard S, An H, Goyal RK, et al. Use of oral anticoagulants among individuals with cancer and atrial fibrillation in the United States, 2010-2016. *Pharmacotherapy*. 2022;42(5):375-386. doi:10.1002/phar.2679

Shapiro SB, Yin H, Yu OHY, Rej S, Suissa S, Azoulay L. Glucagon-like peptide-1 receptor agonists and risk of suicidality among patients with type 2 diabetes: active comparator, new user cohort study. *BMJ*. 2025;388:e080679. Published 2025 Feb 26. doi:10.1136/bmj-2024-080679

Study Designs

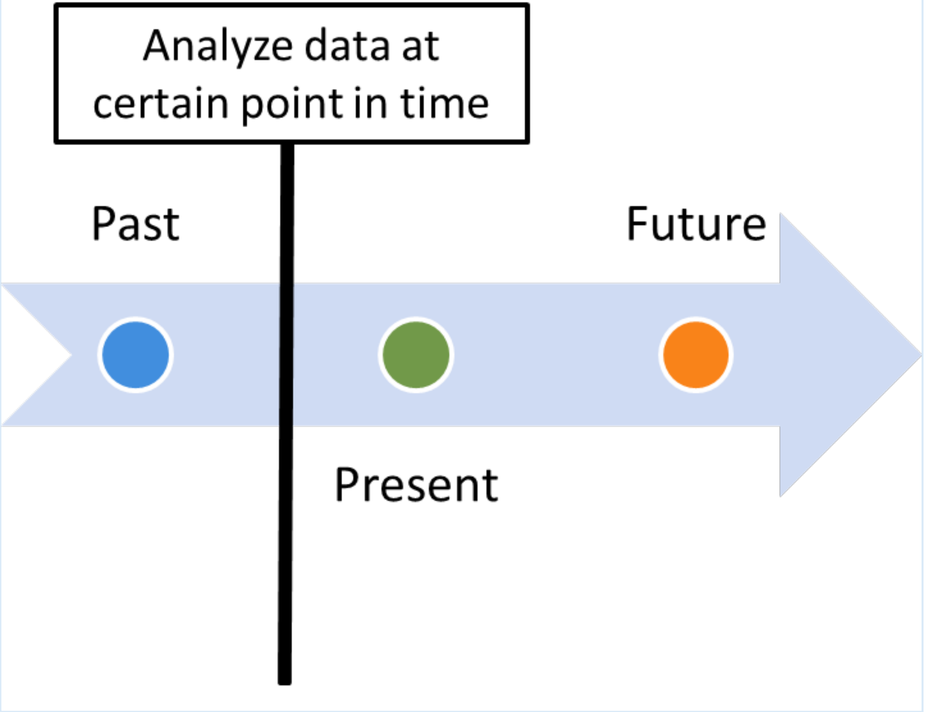
What Are Study Designs?

- Scientific plan to answer a question
  - Similar to a “blueprint” before construction begins

- Good study design is important to conduct valid pharmacoepidemiology (p’epi) research
  - Study design is more important than analysis
  - Poorly designed study cannot be saved by fancy analysis

- Several study designs
  - High-level overview of 3 study designs
    1. Cross-sectional study
    2. Cohort study
    3. Case–control study

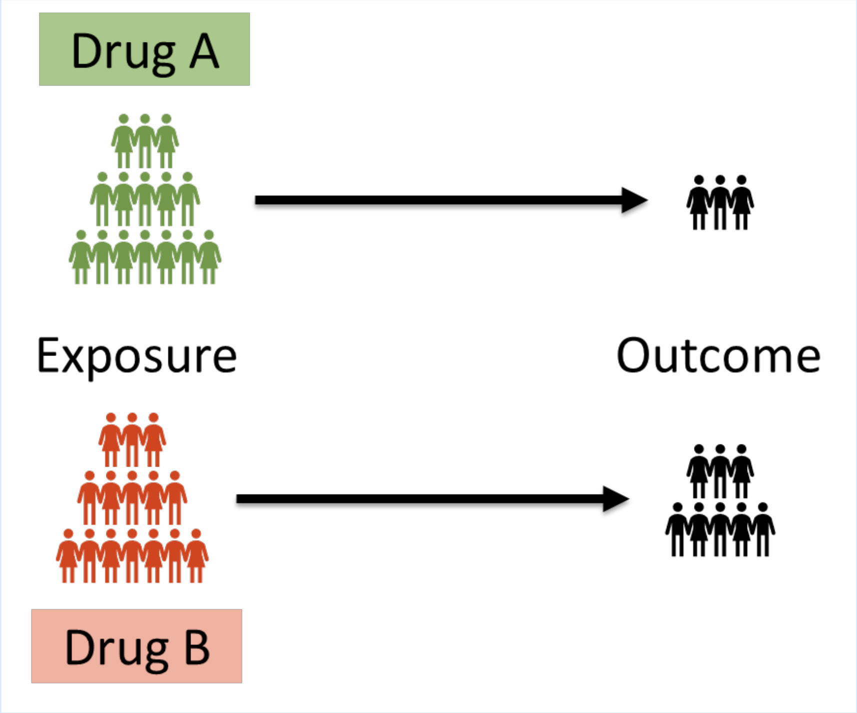
Cross-Sectional Study



- Snapshot in time
  - Time unit: day, week, month, year
- Describe study population or outcomes in a single point in time
- For example:
  - What percentage of people have diabetes in 2024?
  - What percentage of people are using drugs for diabetes in 2024?

Image source: Hemalkumar B. Mehta.

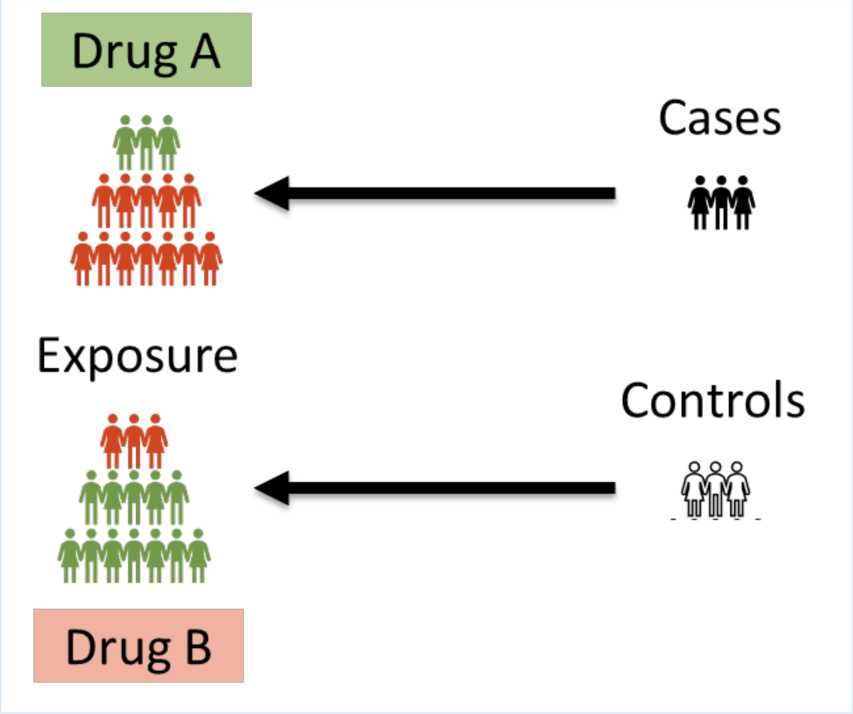
Cohort Study



- Follow people forward in time
- Identify people who are exposed to two different exposures (e.g., drug A vs. drug B) and follow them over time to observe outcomes
- For example:
  - Is drug A better than drug B to reduce the risk of mortality?
  - Does drug A lead to more adverse events than drug B?

Image source: Hemalkumar B. Mehta.
Source: Pottegård A. Core concepts in pharmacoepidemiology: Fundamentals of the cohort and case-control study designs. *Pharmacoepidemiol Drug Saf*. 2022;31(8):817-826. doi:10.1002/pds.5482

Case–Control Study



- Follow people backward in time
- Identify people with outcome and without outcome
  - Follow them backward in time to see who is exposed to something (e.g., drug A vs. drug B)
- For example:
  - Does drug A lead to more adverse events than drug B?
  - Useful for rare outcomes

Image source: Hemalkumar B. Mehta.

Source: Pottegård A. Core concepts in pharmacoepidemiology: Fundamentals of the cohort and case-control study designs. *Pharmacoepidemiol Drug Saf*. 2022;31(8):817-826. doi:10.1002/pds.5482

Statistical Analysis

Produced by the Center for Teaching and Learning at the Johns Hopkins Bloomberg School of Public Health.

Need for Statistical Analysis in Real-World Studies

- In randomized controlled trials, experimental design and randomization help lessen the need for advanced statistical analysis methods

- In real-world studies, there are variety of reasons …
  - Why some people receive treatment versus some don’t
  - Why some will receive drug A versus drug B

- Need to account for such reasons is important
  - First, account at study design level
  - Second, account by conducting appropriate statistical analysis

Some Statistical Analysis Methods

- Descriptive analysis
  - Characterize the study population
  - Describe drug use
- Regression analysis
  - Linear regression for continuous outcomes
  - Logistic regression for binary outcomes
  - Poisson regression for count outcomes
  - Cox regression for survival outcomes

- Propensity score analysis
  - Matching
  - Stratification
  - Weighting
  - Covariate adjustment
- Advanced causal inference methods
  - Instrumental variable analysis
  - Marginal structural models
  - G-computation

Software

Produced by the Center for Teaching and Learning at the Johns Hopkins Bloomberg School of Public Health.

Commonly Used Software

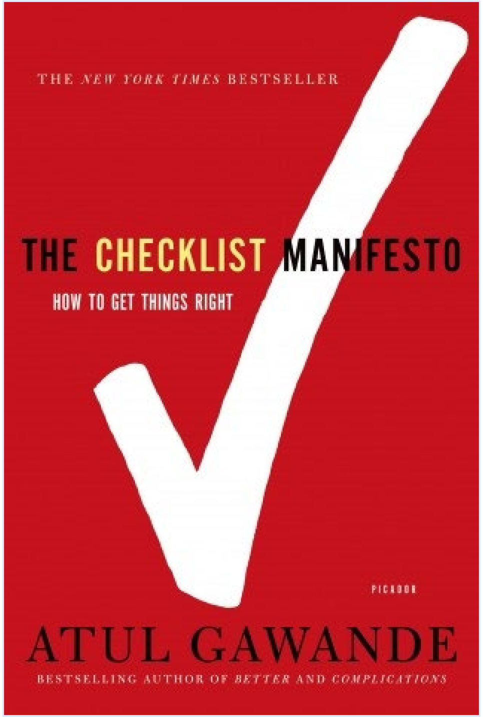
Need software to analyze data in pharmacoepidemiologic research

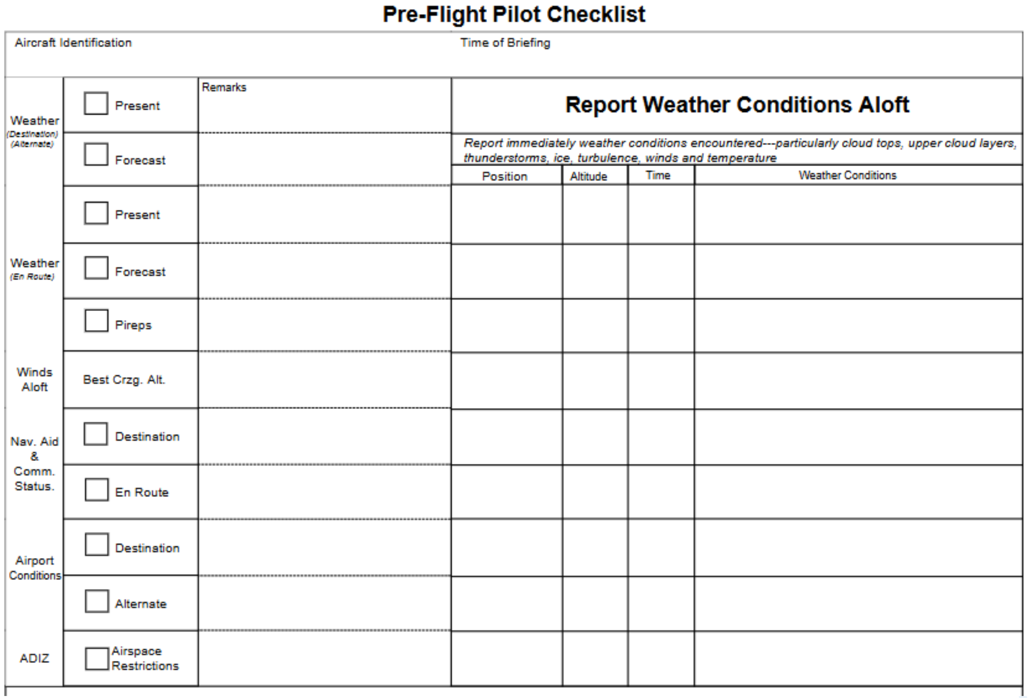
Software	Website	Cost
SAS	https://www.sas.com	Free for student
R	https://www.r-project.org/	Free
STATA	https://www.stata.com/	Discounted price for education
SPSS	https://www.ibm.com/products/spss-statistics/	Discounted price for education
Microsoft Excel	https://www.microsoft.com/en-us/microsoft-365/excel	Discounted price for students
SQL	Multiple implementations of the language	Free
Python	https://www.python.org/	Free

Checklist and Protocols

Produced by the Center for Teaching and Learning at the Johns Hopkins Bloomberg School of Public Health.

Checklist—It Works!


THE NEW YORK TIMES BESTSELLER
THE CHECKLIST MANIFESTO
HOW TO GET THINGS RIGHT
PICADOR
ATUL GAWANDE
BESTSELLING AUTHOR OF BETTER AND COMPLICATIONS



Pre-Flight Pilot Checklist

Aircraft Identification

Time of Briefing

		Remarks
Weather (Destination) (Alternate)	☐ Present	
	☐ Forecast	
Weather (En Route)	☐ Present	
	☐ Forecast	
	☐ Pireps	
Winds Aloft	Best Crzg. Alt.	
Nav. Aid & Comm. Status.	☐ Destination	
	☐ En Route	
Airport Conditions	☐ Destination	
	☐ Alternate	
ADIZ	☐ Airspace Restrictions	

Report Weather Conditions Aloft

Report immediately weather conditions encountered---particularly cloud tops, upper cloud layers, thunderstorms, ice, turbulence, winds and temperature

Position	Altitude	Time	Weather Conditions

Sources:
https://atulgawande.com/book/the-checklist-manifesto/
https://www.faa.gov/documentLibrary/media/Form/7233-4_02-07-24.pdf

Need for Protocols in Pharmacoepidemiology

- Different guidelines and templates exists for the best conduct of real-world studies

- Overarching goal is to have transparency, reproducibility, and clear communication

- Equator network has several guidelines to enhance the quality and transparency of heath research (https://www.equator-network.org/)

- Two examples:
  - STaRT-RWE—Structured template for planning and reporting on the implementation of real-world evidence studies
  - HARPER—HARmonized Protocol Template to Enhance Reproducibility of hypothesis evaluating real-world evidence studies on treatment effects

STaRT-RWE

- A public–private consortium developed a structured template for planning and reporting on the implementation of real-world evidence studies of the safety and effectiveness of treatments

- Appendix has a template and also provides some examples for real-world studies

Table of contents:

1. Administrative information
   - Title, objective, protocol registration, ethics
2. Version history
3. Design diagram
4. Summary of analytics study population
5. Analysis specification
6. Sensitivity analyses
7. Attrition table
8. Power and sample size calculations
9. Glossary of terminology
10. Abbreviations

Source: 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. Published 2021 Jan 12. doi:10.1136/bmj.m4856

HARPER

- International Society for Pharmacoepidemiology (ISPE) and International Society for Pharmacoeconomics and Outcomes Research (ISPOR) convened a joint task force to create a template for real-world studies

- Builds on existing efforts

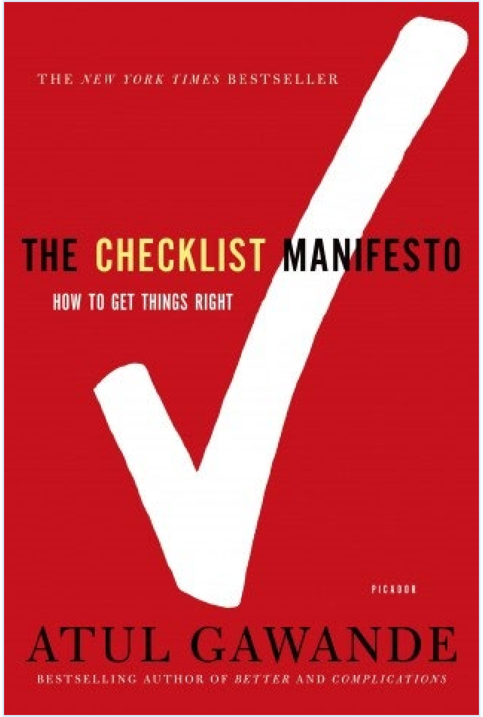
- Sample templates available on Github website that incorporate key components

Table of contents

1. Title page
2. Abstract
3. Amendments and updates
4. Timeline
5. Rationale and background
6. Research question and objectives
7. Research methods
8. Limitations
9. Human subject research
10. Reporting of adverse events
11. References
12. Appendices

Sources:
Wang SV, Pottegård A, Crown W, et al. HARmonized Protocol Template to Enhance Reproducibility of hypothesis evaluating real-world evidence studies on treatment effects: A good practices report of a joint ISPE/ISPOR task force. *Pharmacoepidemiol Drug Saf*. 2023;32(1):44-55. doi:10.1002/pds.5507
https://drugepi.gitlab-pages.partners.org/harper/

Overall Idea


THE NEW YORK TIMES BESTSELLER
THE CHECKLIST MANIFESTO
HOW TO GET THINGS RIGHT
PICADOR
ATUL GAWANDE
BESTSELLING AUTHOR OF BETTER AND COMPLICATIONS

- Guideline or checklist can help in several ways by:
  - Improving transparency
  - Improving clear communication
  - Improving credibility
  - Improving reproducibility
  - Reducing confusion
  - Reducing errors/mistakes
  - Increasing acceptance by regulators, decision makers, and payers

Source: https://atulgawande.com/book/the-checklist-manifesto/

Summary

Produced by the Center for Teaching and Learning at the Johns Hopkins Bloomberg School of Public Health.

Key Points

- **Randomization vs. the real world:** Real-world studies can complement randomized controlled trials for evidence generation
- **Data sources:** Variety of data sources are available; pick one that fits the need
- **Study designs:** Good study design is a prerequisite; cross-sectional, cohort, and case–control
- **Statistical analysis:** Need appropriate analysis methods based on the research question
- **Software:** Several options to choose from—SAS and R
- **Checklist and protocols:** Can improve transparency, reproducibility, and clear communication