EPID 765
Pharmacoepidemiology

Lesson 2

Sources of Data for Pharmacoepidemiology

Presenter: Mitch Conover
Date: January 15, 2019

1

Outline

- Introduction to automated databases for PE
- Types of automated databases
  - Administrative
  - Medical registry
  - Clinical
  - Electronic health record
- Examples of automated databases
- PE outside of automated databases
- Where are we headed?

2

Acknowledgements

- Slides adapted from Til Sturmer & Michele Jonsson Funk
- Thank you to Tian (course T.A.) for helping to coordinate today's lecture

3

PE Data Considerations

- Need for timely results
  - Public / individual health
  - Regulatory / commercial
- Use of specific drug limited by
  - Indication
  - Recency of market introduction
  - Competitors (choice)
- Resulting requirements
  - Already collected ("retrospective")
  - Large N (often > 1M)
    - Note: All of Us (Precision Medicine) Initiative

4

Advantages of Automated Databases for Pharmacoepidemiology

- Data already collected (timely!)
- Large N (sometimes huge!)
- Inexpensive (well … kind of)
- Often +/- population-based
  - Little potential for selection bias
- Clear timeline
  - Prospectively collected
  - No recall/interviewer bias
- Efficient (secondary use)

5

Ideal Database

- Large
- Timely (i.e., up to date)
- Structured (e.g., dx codes)
- Continuity
  - Individual observations
  - Calendar time
- Linkage on unique identifier
- Potential for
  - Chart review
  - Access to patients
- Accessible
  - Without delay
  - Over prolonged periods
  - To everyone (?)
- Representative (?)

6

Ideal Database: Data Elements

- Prescription drugs
- Over-the-counter (OTC) drugs
- Outpatient, inpatient, emergency care
- Mental health care
- Indication for treatment, e.g.,
  - Diagnoses
  - Laboratory
  - Radiographic
  - Function (RR, ejection fraction)
- Other determinants of treatment and outcome
  - Insurance plan
  - Prescriber
  - SES
  - BMI, smoking, diet, exercise, frailty
- Cause-specific mortality
- Patient reported outcomes (PROs, e.g., QOL)

7

7

Real Databases

- None is ideal
- Trade-off between
  - Advantages
  - Disadvantages
- For specific research question
- Think out of the box
- Secondary data: ”for another purpose”
- Consider possibilities of linkage
  - Deterministic vs. probabilistic
  - Vertical vs. horizontal

8

8

Administrative Databases

- Medical care data
- Not collected for research/patient care
- Often generated for reimbursement
- Representing medical transactions
  - Generally good for
    - High cost, e.g., biologics, chemotherapies, surgery
    - Acute events (e.g., hip fracture)
  - Less accurate for
    - Low cost, e.g., generic drugs
    - Chronic diseases
      - Low sensitivity, e.g., hypertension
      - Low specificity, e.g., rule-out diabetes

9

9

Healthcare Claims Databases

- Based on fee for service system
- Every financial transaction results in “paper” trail
- Highly structured (no free text)
- Benefits from auditing of financial transactions
  - Fraud checking (stiff penalties)
- Payor imposes minimal requirements
  - Little missing data
- Usually obtained from payor
- Data use agreement
- Personal identifiers stripped
- No informed consent

10

10

Healthcare Claims Databases

- Advantages:
  - Large, timely, longitudinal
  - “Clean”
  - Data checks (built into the adjudication process)
  - Financial incentive to report
  - Essentially no additional cost
  - Capture care across many providers/facilities
- Disadvantages
  - No data on lifestyle
    (no data on BMI does not imply biased!)
  - Codes do not guarantee exposure/outcome
  - US: under 65 limited by changing payors
  - Chart review/contact difficult/impossible

11

11

Example Claims Data: Medicare

- All individuals ≥65 years and those <65 with particular disabilities qualify for federally funded health insurance
- Population-based (almost)
- Research is conducted on **administrative/ billing data**
  - Often limited to individuals with continuous enrollment and “fee-for-service” coverage
- Prescription drug data are captured through Medicare Part D (active since 1/1/2006)
- Additional drug data can be obtained using State pharmacy assistance plans (e.g., NJ, PE)

12

12

Medical Registry Databases

- Databases often collected as part of a government or regulatory mandate
- Focus on specific disease/procedure/treatment
  - Involve additional abstraction of information (e.g., tumor histology, nodal involvement, etc)
  - Can cost a lot $$ to maintain
  - Delay in reporting
- Registration may be
  - Legally binding
  - Prerequisite for treatment
- E.g., cancer registry, STIs, UK biologic registry

13

13

Example Medical Registry: SEER

- Surveillance Epidemiology and End Results
- Since 1974
- One of NCI's most important data collection and dissemination activities
- System of population-based cancer registries strategically located across the US
- Monitor cancer trends
- Provide timely, accurate, and continuous data on
  - Cancer incidence
  - Extent of disease at diagnosis
  - Therapy
  - Patient survival

14

14

Electronic Health Records Databases

- Ideal: complete lifetime medical record
- Linked across multiple health care providers
- Owned by patient?
- Instantly available to all healthcare providers
- Unfortunately in US: not even close
  - Different systems
  - Ownership issues
  - Legal issues
- CPRD, Kaiser, Regenstrief etc. come close
- In US very dynamic but not uniform!

15

15

Electronic Health Record (EHR)

- Data collected for patient care purposes (i.e., patient medical record)
- May be incomplete due to delivery of patient care in different settings
  - E.g., information collected on an inpatient basis may not be recorded in the same place as outpatient data
- EHR preferred over EMR (more inclusive)
- EHR also used to indicate continuity
- Validation of diagnoses moot as data represent gold standard(?)

16

16

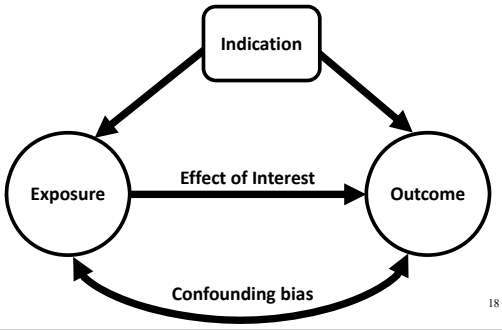
Disease vs. diagnosis

- Is EHR a perfect indicator of disease onset? NO.
- [In some settings] EHR may be a perfect indicator for *diagnosis* of disease
  - May be the relevant confounder in your study
- Some diseases: no precise point of onset
- EHR marks when patients seek medical attention
  - May happen late in disease progression (e.g. cancer, diabetes)

17

17

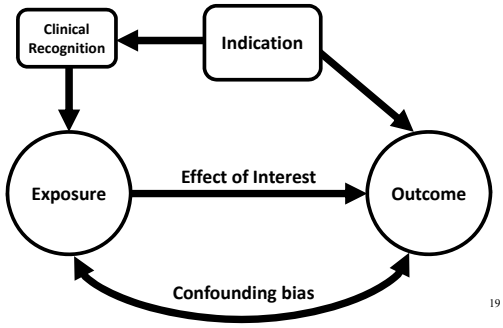
E.g. Confounding by indication



18

18

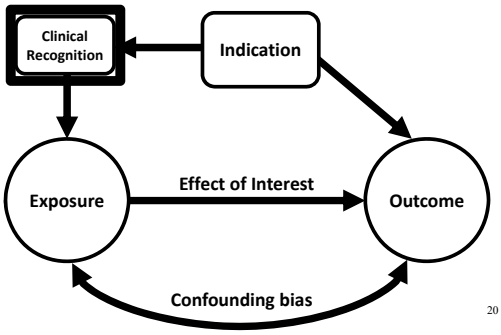
E.g. Confounding by indication



19

19

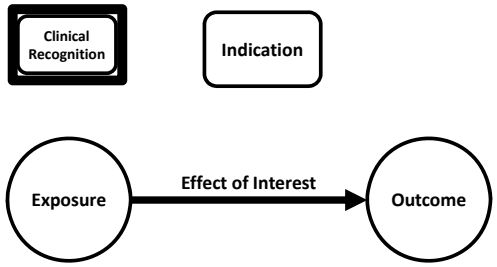
E.g. Confounding by indication



20

20

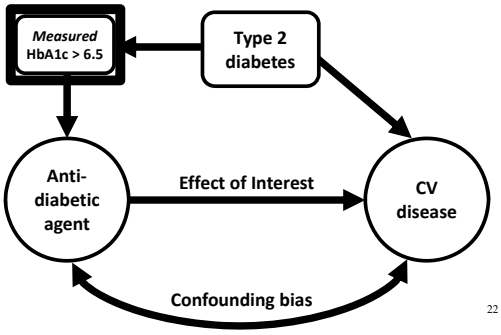
E.g. Confounding by indication



21

21

E.g. Confounding by indication



22

22

8 domains of EMR misclassification

1. EMR data reflect only the health services and medications delivered within the specific health care setting that con-tributes to the EMR system [1, 2••, 3, 4••, 5, 6]. This leads to both left and right censoring, and uncertainty regarding the person-time at risk. This is particularly problematic in inpatient EMRs.
2. Prescription records in an ambulatory EMR reflect clinician orders for medications, which may not be filled or consumed by the patient [7–9].
3. In EMR studies, defining treatment episodes/treatment duration/cumulative exposure is complex and requires many decisions which have unpredictable influence on exposure misclassification [10, 11••, 12].

Young JC, Conover MM, Jonsson Funk M. Measurement Error and Misclassification in Electronic Medical Records: Methods to Mitigate Bias. Current Epidemiology Reports (2018) 5:343–356

23

23

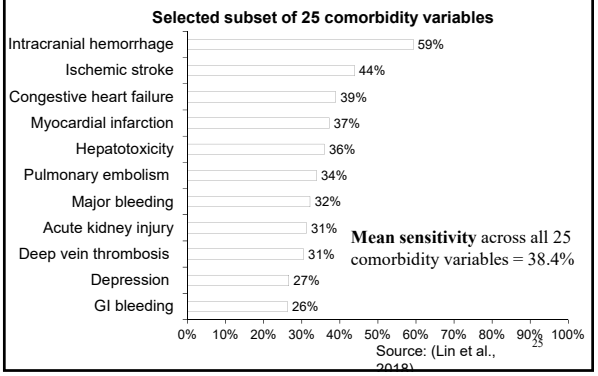
8 domains of EMR misclassification

4. Automated data entry in EMR systems may forward-propagate erroneous data and/or carry forward information that is no longer clinically relevant [13–15].
5. Recent advances in natural language processing (NLP),which automate extraction of information from unstructured data, may introduce systematic errors [16–19].
6. Performance of EMR-based clinical prediction algorithms may vary widely between different health systems [20•].
7. Temporal changes in the recording of EMR data elements may produce systematic differences in classification and/or missingness over time [21].
8. Horizontal linkage of populations captured by different EMR systems produces systematic differences in classification and/or missingness between the linked populations[22].

24

24

Sensitivity of variables defined in EMR vs. EMR + claims

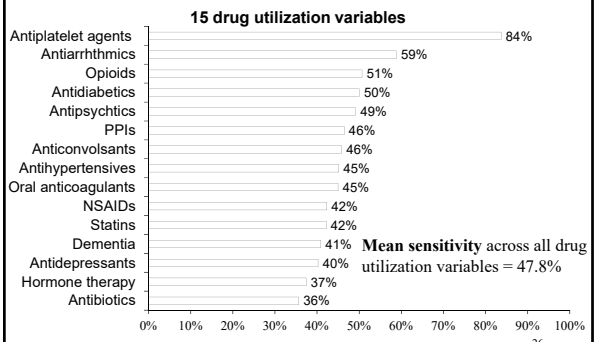


Source: (Lin et al., 2018)

25

25

Sensitivity of variables defined in EMR vs. EMR + claims

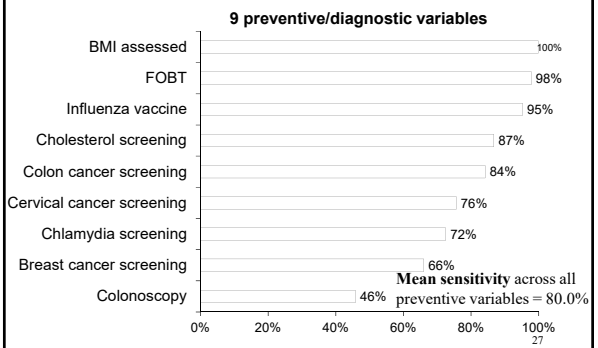


Source: (Lin et al., 2018)

26

26

Sensitivity of variables defined in EMR vs. EMR + claims



Source: (Heintzman et al., 2014)

27

27

Example EHR: Clinical Practice Research Datalink (CPRD)

- Formerly known as VAMP, GPRD
- Medical records >10 million patients
- Data collected by general practitioner (GP)
- Used as chart (as part of patient care)
- GP is gatekeeper in UK: good continuity
- Possible to access medical records ($$)
- Drugs prescribed (not: dispensed!)
- Linkage to hospital data in England only

28

28

Example EHR: Kaiser Permanente

- Vertically integrated health provider covering 8.2 million people from 8 geographic regions
- Well-defined population
- Includes all primary, specialty and most emergency and hospital care
- Administrative and clinical data
- Part of HMO-RN and CRN
- Each individual is provided a unique Kaiser identification number
  - Follow-up over time
  - Linkage between databases
- Cave: relatively strict formulary!

29

29

Example EHR: Regenstrief

- Informatics and healthcare research organization
- Established 1969 by Sam Regenstrief
- Indiana University - Purdue University
- Regenstrief Medical Records System (RMRS)
- Nation's only citywide electronic medical records system which currently allows emergency department physicians, with the patient's permission, to view as a single virtual record all previous care at any of 18 participating hospitals

30

30

Example EHR: Carolina Data Warehouse

- UNC Health Care System wide
- Enhancement quality of care & clinical research
- Central repository
  - Clinical, research, administrative data
  - Billing, insurance, diagnosis, and medication
- Data since Jul-2004 refreshed every 24-48 hrs
- Research portal offers a Cohort Discovery service as a pre-research step
  - Basic queries w/ i2b2 (brief training for access)
- Accessible for everyone within UNC (via DUA)
- EPIC system since Aril-2014

31

31

Examples Population Based Data: Scandinavian Databases

- When an entire country is a cohort
  (Lone Frank. Science 2000;287 (5462):2398-2399)
- 6.5M
- Population-based
- Universal healthcare
- Unique identifier
  - Constant over lifetime
  - Includes checknum (100% linkage ☺)
- Societal agreement to use data for research
  - Including genetic data (dynamic opt out, N~300!)

32

32

Other Data: Clinical Databases

- Data specifically collected for research and/or quality improvement purposes
- Specific disease/procedure/treatment
- Contain structured information not usually captured in EHR or claims
  - E.g., Performance scores, smoking status, alcohol intake, disease specific scores, etc.
- Increasingly conceptualized as add-on to EHR

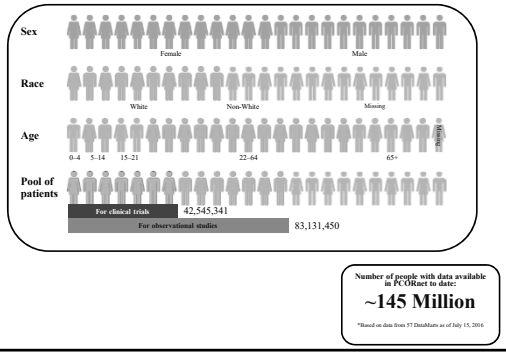
33

33

Data networks: Sentinel (claims-based)

34

Data Networks: PCORnet (EMR-based)



35

Other Data: Ad Hoc Studies

- RCTs
- Cohort studies
- Case-control studies
- Cross-sectional studies
- Limited by N and collection intervals
- Check how drug use was assessed
  - Interview (+/- cabinet; probing?) vs. self report
  - Period (e.g., last month vs. since last interview)
- Drug registries
  - E.g., antiretrovirals, biologics
  - Often limited by lack of comparator

36

Where are we headed? Data linkage

- Electronic Medical Records
  - Single system, multi-system
- Insurance claims
  - Medicare, Medicaid, Commercial
- Registries Primary data
  - Patient reported outcomes
  - Research-specific assessments
- Vital statistics (birth, death)
- Lab data, imaging, pathology
- Genomic data
- Aggregate data from geographical units
  - Air pollution, water quality
  - Weather (heat waves)
  - SES
- New tech
  - Wearables, smart phone apps
  - Skin patches, smart pills/bottles
- Social media, internet searches, purchases (pregnancy tests)

37

Data linkage examples

- Add data to cohort studies
  - Easy to get informed consent
  - E.g., ARIC, WHI, Rotterdam
- Add claims data to registries
  - E.g., SEER-Medicare
- Internal validation studies
  - Add additional information for a sub-group
  - E.g., Medicare Current Beneficiary Survey (MCBS)
  - Chart review
  - Cause of death data
- Add disease registries to EHR data
  - E.g., cancer registry

38

Active Linkages to UNC's CDW

- NC death certificate data
- Blue Cross Blue Shield of NC
  - Individual market + state employees
  - 30% of UNCHS patients
- Medicare
  - 2016: All fee-for-service patients in 2016
  - 2017: early 2019
  - 20% of age 65+ with Part D coverage for 2007-2016 (for some years 100% of CDW)

39

The Future of Data for PE

- Dependent on
  - Safeguards against misuse
    - Privacy (esp. medical chart data)
    - Bad science (difficult to define)
  - Acceptance of research(!)
    - Society
    - Stakeholders
  - Generational contract
    - Individual consent impossible/defies purpose
    - I benefit from data of prior patients and future patients benefit from mine
    - Utilitarian principle OK given low risk!

40

40

External links

- Carolina Data Warehouse: Information about i2b2 system and training
  - https://tracs.unc.edu/index.php/services/informatics-and-data-science/i2b2
- A list of UNC Pharmacoepidemiology Data Resources and a (partial) list of the students who work with them:
  - http://goo.gl/GmRdFY
- UNC Digital Health Initiative
  - https://hsl.lib.unc.edu/digital-health/
- Young JC, Conover MM, Jonsson Funk M. Measurement Error and Misclassification in Electronic Medical Records: Methods to Mitigate Bias. Current Epidemiology Reports (2018) 5:343–356.
  - Shareable link: https://rdcu.be/6dmQ

41

41