Electronic Health Records and Big Data Analytics

Christopher Meaney

Biostatistician
Department of Family and Community Medicine
University of Toronto
&
PhD Candidate
Dalla Lana School of Public Health
Division of Biostatistics
University of Toronto

October 27, 2020

Medical record keeping has been with us since the ancient times.

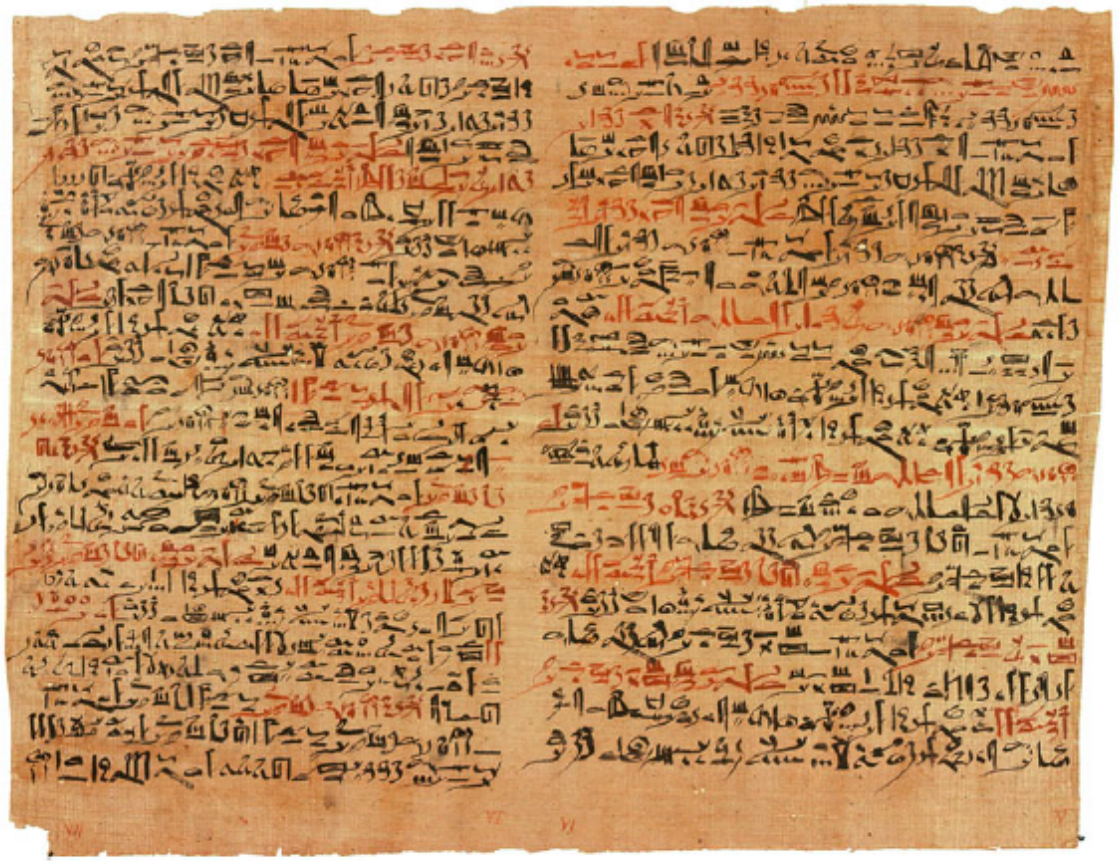
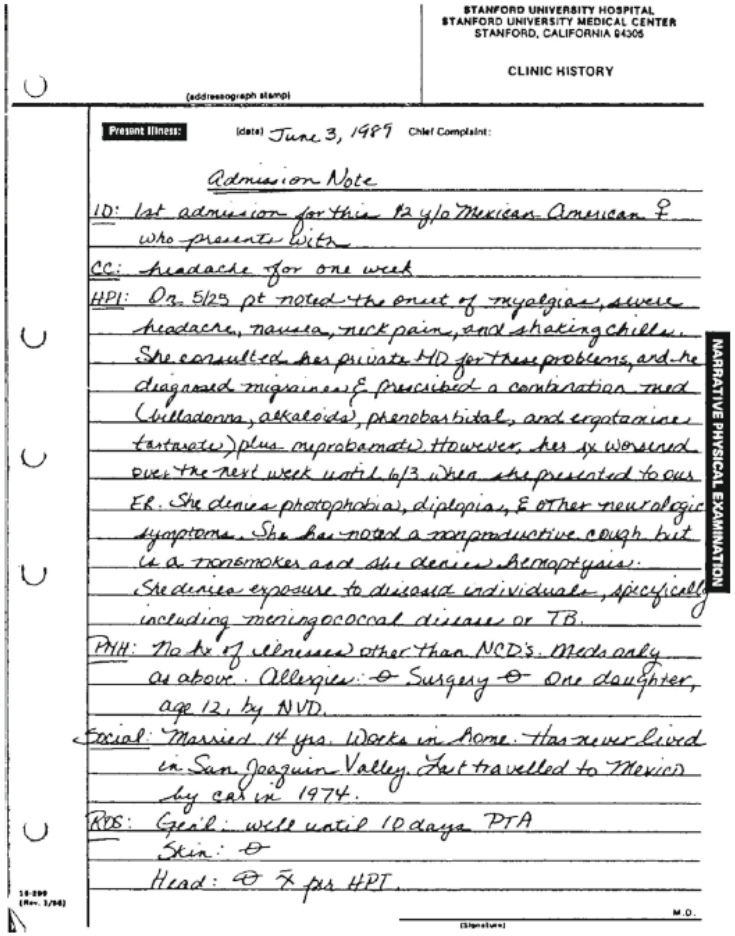


Fig. 2.1 Part of the Edwin Smith Papyrus describing in Egyptian hieratic script (a cursive hieroglyph writing) different surgery cases from 1600 BC (Published in Wikipedia)

Dalianis, H. (2018). Clinical Text Mining. Pg. 9.

Medical records centuries later. Serving different purposes over time.



STANFORD UNIVERSITY HOSPITAL
STANFORD UNIVERSITY MEDICAL CENTER
STANFORD, CALIFORNIA 94305

CLINIC HISTORY

(addressograph stamp)

Present Illness: (date) June 3, 1989 Chief Complaint:

Admission Note

ID: 1st admission for this 12 y/o Mexican-American ♀ who presents with

CC: headache for one week

HPI: On 5/25 pt noted the onset of myalgias, severe headache, nausea, neck pain, and shaking chills. She consulted her private MD for these problems, and he diagnosed migraines & prescribed a combination med (belladonna alkaloids), phenobarbital, and ergotamines tartrate) plus suprobamate). However, her sx worsened over the next week until 6/3 when she presented to our ER. She denies photophobia, diplopia, & other neurologic symptoms. She has noted a nonproductive cough but is a nonsmoker and she denies hemoptysis. She denies exposure to diseased individuals, specifically including meningococcal disease or TB.

PMH: No hx of illnesses other than NCD's. Meds only as above. Allergies: ⊖ Surgery ⊖ One daughter, age 12, by NVD.

Social: Married 14 yrs. Works in home. Has never lived in San Joaquin Valley. Pt travelled to Mexico by car in 1974.

ROS: Gen'l: well until 10 days PTA
Skin: ⊖
Head: ⊖ ✗ per HPT.

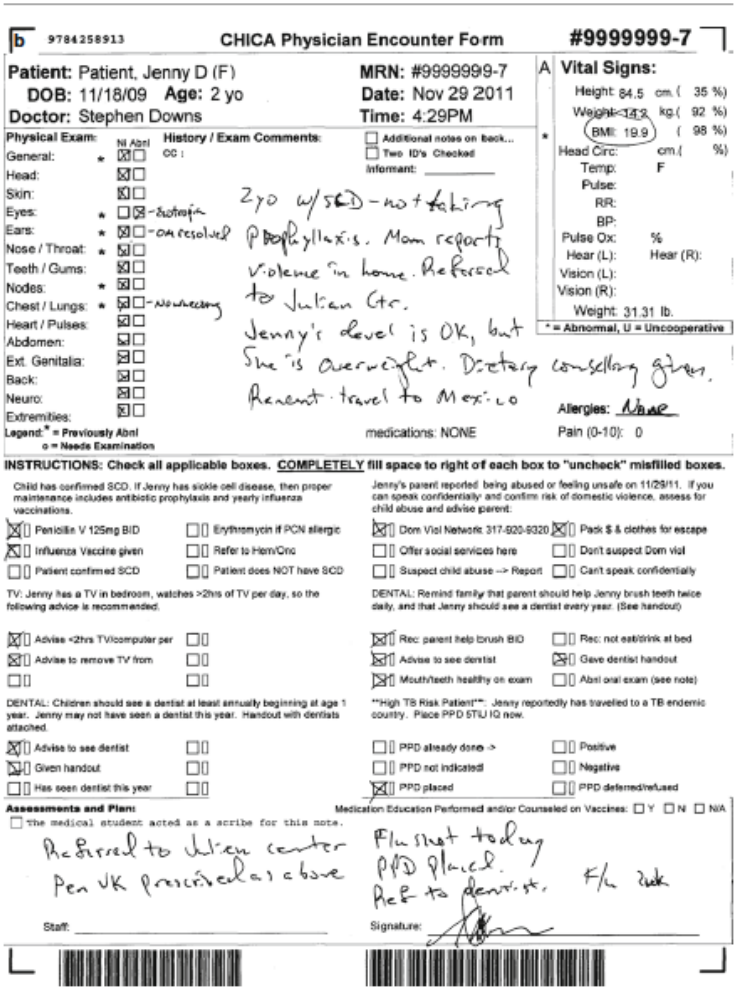
M.D.
(Signature)

NARRATIVE PHYSICAL EXAMINATION

Fig. 2.1 Much of the information gathered during a physician–patient encounter is written in the medical record

Shortliffe, E., Camino, J. (2014). Biomedical Informatics. p. 42.

The medical record slowly integrated structured data elements. [1]



CHICA Physician Encounter Form #9999999-7

Patient: Patient, Jenny D (F) MRN: #9999999-7
DOB: 11/18/09 Age: 2 yo Date: Nov 29 2011
Doctor: Stephen Downs Time: 4:29PM

Vital Signs:
Height 84.5 cm. (35 %)
Weight 14.2 kg.(92 %)
BMI 19.9 (98 %)
Head Circ: cm.(%)
Temp: F
Pulse:
RR:
BP:
Pulse Ox: %
Hear (L): Hear (R):
Vision (L):
Vision (R):
Weight: 31.31 lb.
* = Abnormal, U = Uncooperative

Physical Exam: (Ni Abnl) — General, Head, Skin, Eyes, Ears, Nose / Throat, Teeth / Gums, Nodes, Chest / Lungs, Heart / Pulses, Abdomen, Ext. Genitalia, Back, Neuro, Extremities
Legend: * = Previously Abnl; o = Needs Examination

History / Exam Comments:
CC:
Additional notes on back...
Two ID's Checked
Informant:

2yo w/SCD - not taking Prophyllaxis. Mom reports violence in home. Referred to Julian Ctr. Jenny's devel is OK, but she is overweight. Dietary counselling given. Recent travel to Mexico

Allergies: None
medications: NONE
Pain (0-10): 0

INSTRUCTIONS: Check all applicable boxes. COMPLETELY fill space to right of each box to "uncheck" misfilled boxes.

Child has confirmed SCD. If Jenny has sickle cell disease, then proper maintenance includes antibiotic prophylaxis and yearly influenza vaccinations.
- [x] Penicillin V 125mg BID
- [x] Influenza Vaccine given
- [ ] Patient confirmed SCD
- [ ] Erythromycin if PCN allergic
- [ ] Refer to Hem/Onc
- [ ] Patient does NOT have SCD

Jenny's parent reported being abused or feeling unsafe on 11/29/11. If you can speak confidentially and confirm risk of domestic violence, assess for child abuse and advise parent:
- [x] Dom Viol Network: 317-920-9320
- [x] Pack $ & clothes for escape
- [ ] Offer social services here
- [ ] Don't suspect Dom viol
- [ ] Suspect child abuse --> Report
- [ ] Can't speak confidentially

TV: Jenny has a TV in bedroom, watches >2hrs of TV per day, so the following advice is recommended:
- [x] Advise <2hrs TV/computer per
- [x] Advise to remove TV from
- [ ]

DENTAL: Remind family that parent should help Jenny brush teeth twice daily, and that Jenny should see a dentist every year. (See handout)
- [x] Rec: parent help brush BID
- [x] Advise to see dentist
- [x] Mouth/teeth healthy on exam
- [ ] Rec: not eat/drink at bed
- [x] Gave dentist handout
- [ ] Abnl oral exam (see note)

DENTAL: Children should see a dentist at least annually beginning at age 1 year. Jenny may not have seen a dentist this year. Handout with dentists attached.
- [x] Advise to see dentist
- [x] Given handout
- [ ] Has seen dentist this year

High TB Risk Patient: Jenny reportedly has travelled to a TB endemic country. Place PPD 5TU IQ now.
- [ ] PPD already done ->
- [ ] PPD not indicated
- [x] PPD placed
- [ ] Positive
- [ ] Negative
- [ ] PPD deferred/refused

Assessments and Plan:
- [ ] The medical student acted as a scribe for this note.

Referred to Julian center
Pen VK prescribed as above

Flu shot today
PPD placed.
Ref to dentist. F/u 2wk

Medication Education Performed and/or Counseled on Vaccines: [ ] Y [ ] N [ ] N/A

Staff: _____ Signature: _____

[1] Shortliffe, E., Camino, J. (2014). Biomedical Informatics. p. 408.

Medical record keeping enters the filing cabinet era!

Transition to Computer-Based EHRs...Digital Health IT Solutions.

What Information Exist in the EHR and What is it Used For?

- Heterogeneous clinical data structures constitute modern EHR systems.
- Data can inform patient portals, and patient-EHR interfaces (in some jurisdictions).
- Patient tech may populate EHR (heart rate; pedometer; hours sleep; hours screen time).
- Communication between clinical and business units (e.g. billing/finance, HR, etc.).
- Uni-directional flow of information/data to research? (feedback, QI)?

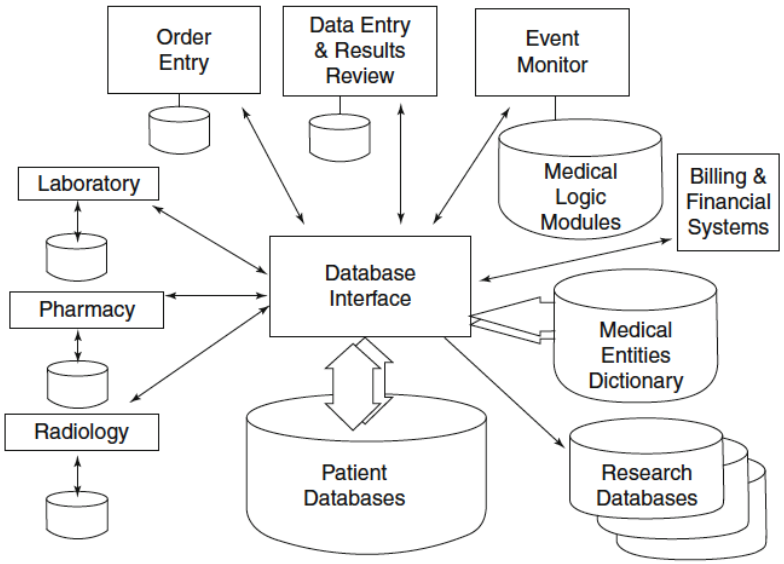
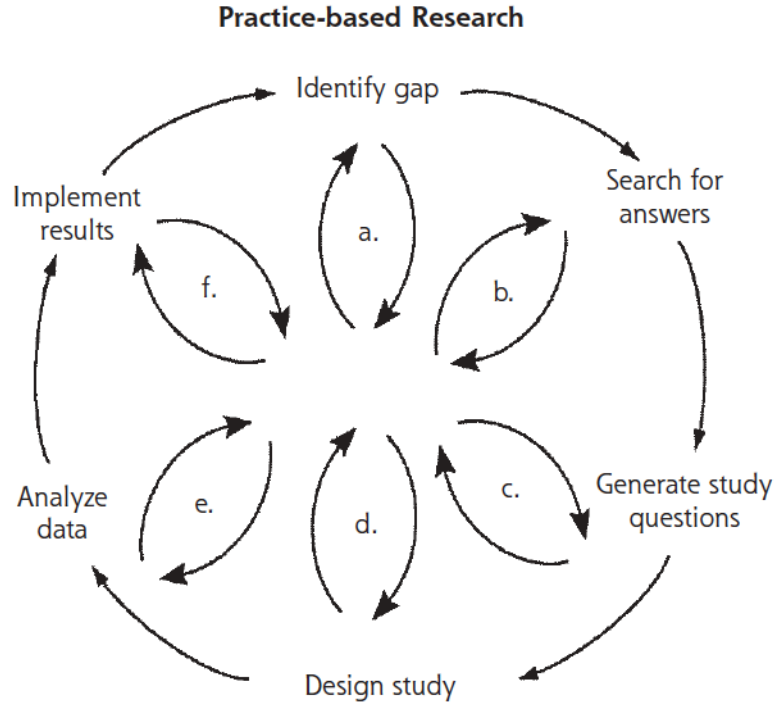


Fig. 12.2 A block diagram of multiple-source-data systems that contribute patient data, which ultimately reside in a computerized patient record (CPR). The database interface, commonly called an interface engine, may perform a number of functions. It may simply be a router of information to the central database. Alternatively, it may provide more intelligent filtering, translating, and alerting functions, as it does at Columbia University Medical Center (Source: Courtesy of Columbia University Medical Center, New York)

Shortliffe, E., Camino, J. (2014). Biomedical Informatics. p. 396.

PBRNs for Improving Health/Healthcare via Embedded Research and QI

Figure 1. Diagram of practice-based research involving clinicians in research steps.



Note: Diagram constructed by Nutting at the Primary Care Methods Conference, San Antonio, Texas, 2000.

Mold et al. (2005). PBRNs at Interface Research at QI. Annals Fam Med.

EHR Research Networks: Foundational Infrastructure

- Governance. Leadership. Buy-in/Engagement.
- Funding/resources.
- Data Stewards/Custodians. Data Partnerships. Data Sharing.
- Consent. Confidentiality. Data ownership? Ethics, trust.
- Information Privacy. Health Data Regulation/Legislation.
- Information technology. Performance. Design/UX. Security. Threat Risk Assessments.
- ETL processes and database curation/management. Generating documentation.
- HQPs: Clinical, Management/Admin, Research (Epi, Biostats, Informatics, AI, etc.)

Selected Primary Care Research Networks:

- Ontario: UTOPIAN, EON, MUSIC, OPEN, DELPHI, NORTHH. EMRPC.
- Canada: CPCSSN.
- International: UK (CPRD, THIN). Oregon, USA (OHSU-PBRN). Singapore (PCRN).

Fewer Hospital-Based EHR Research Networks in Canada. Why?

- Canada: GEMINI
- USA: MIMIC-III (Open Access).

Tierney et al. (2007). National Survey of PBRNs. Annals Fam Med.

Davis et al. (2015). Char and Lessons Learned PBRNs in USA. J Hlthcare Leadership.

Why Should/Does the Health Care System Support/Adopt EHRs?

- (A belief that...) EHRs improve the health of patient/communities.
- Earlier evidence (pre-2010), suggestion EHRs may not achieve goals?

Background

- In 1991, the Institute of Medicine released a landmark report recommending that electronic health records (EHRs) be implemented in health systems within 10 years[1];
- Almost 20 years later, only a small proportion of health providers have implemented EHRs.[2 3]
- A meta-analysis of diffusion rates of EHRs in the U.S. shows that uptake has slowed in recent years. The study concludes: "EHRs are the future, and resistance is futile; however, current exigencies and uncertainties are slowing, not accelerating adoption."[4]
- The very definition of what comprises an EHR remains in flux e.g. Web-based patient self-entry systems to fully integrated and interoperable systems linking multiple providers.[5 6]
- Empirical evidence on the benefits of EHRs is scarce;[7 8]
- Much of the published evidence on EHRs is conflicting; a broad variety of research disciplines representing widely varying approaches and perspectives contribute to this body of literature and this has been indicated as partly to blame for these discrepancies.[9]

Grimshaw et al; Evidence Summary: Electronic Health Records. LHIN Report.

Some Evidence EHRs Improving Hospital Quality and Patient Satisfaction

Assessing the Impact of Electronic Health Records as an Enabler of Hospital Quality and Patient Satisfaction

Benjamin Jarvis, MS-HSM, Tricia Johnson, PhD, Peter Butler, MHSA, Kathryn O'Shaughnessy, MS-HSM, Francis Fullam, MA, Lac Tran, MSSD, and Richa Gupta, MBBA, MHSA

Abstract

Purpose
To assess the impact of using an advanced electronic health record (EHR) on hospital quality and patient satisfaction.

Method
This retrospective, cross-sectional analysis was conducted in 2012 to evaluate the association between advanced EHR use (Healthcare Information Management Systems Society [HIMSS] Stage 6 or 7 as of December 2012) and estimated process and experience of care scores for hospitals under the Medicare Hospital Value-Based Purchasing Program, using data from the American Hospital Association for 2008 to 2010. Generalized linear regression models were fit to test the association between advanced EHR use with process of care and experience of care, controlling for hospital characteristics. In a second analysis, the models included variables to account for HIMSS stage of advanced EHR use.

Results
The study included 2,988 hospitals, with 248 (8.3%) classified as advanced EHR users (HIMSS Stage 6 or 7). After controlling for hospital characteristics, advanced EHR use was associated with a 4.2-point-higher process of care score ($P < .001$). Hospitals with Stage 7 EHRs had 11.7 points higher process of care scores, but Stage 6 users had scores that were not substantially different from those of nonadvanced users. There was no significant difference in estimated experience of care scores by level of advanced EHR use.

Conclusions
This study evaluated the effectiveness of the U.S. federal government's investment in hospital information technology infrastructure. Results suggest that the most advanced EHRs have the greatest payoff in improving clinical process of care scores, without detrimentally impacting the patient experience.

Jarvis et al. (2013). EHRs and Hospital Quality and Patient Satisfaction. Acad Med.

Some Evidence EHRs Improving Hospital Quality and Patient Satisfaction

EHR Adoption and Hospital Performance: Time-Related Effects

Julia Adler-Milstein, Jordan Everson, and Shoou-Yih D. Lee

Objective. To assess whether, 5 years into the HITECH programs, national data reflect a consistent relationship between EHR adoption and hospital outcomes across three important dimensions of hospital performance.

Data Sources/Study Setting. Secondary data from the American Hospital Association and CMS (Hospital Compare and EHR Incentive Programs) for nonfederal, acute-care hospitals (2009–2012).

Study Design. We examined the relationship between EHR adoption and three hospital outcomes (process adherence, patient satisfaction, efficiency) using ordinary least squares models with hospital fixed effects. Time-related effects were assessed through comparing the impact of EHR adoption pre (2008/2009) versus post (2010/2011) meaningful use and by meaningful use attestation cohort (2011, 2012, 2013, Never). We used a continuous measure of hospital EHR adoption based on the proportion of electronic functions implemented.

Data Collection/Extraction Methods. We created a panel dataset with hospital-year observations.

Principal Findings. Higher levels of EHR adoption were associated with better performance on process adherence (0.147; $p < .001$) and patient satisfaction (0.118; $p < .001$), but not efficiency (0.01; $p = .78$). For all three outcomes, there was a stronger, positive relationship between EHR adoption and performance in 2010/2011 compared to 2008/2009. We found mixed results based on meaningful use attestation cohort.

Conclusions. Performance gains associated with EHR adoption are apparent in more recent years. The large national investment in EHRs appears to be delivering more consistent benefits than indicated by earlier national studies.

Key Words. Health care organizations and systems, hospitals, information technology

Adler et al. (2015). EHR Time Related Effects and Hospital Performance. HSR.

Mixed Evidence with Respect to EHRs on Hard Patient Outcomes?

HEALTH INFORMATION TECHNOLOGY

By Sunny C. Lin, Ashish K. Jha, and Julia Adler-Milstein

Electronic Health Records Associated With Lower Hospital Mortality After Systems Have Time To Mature

ABSTRACT Evidence linking electronic health record (EHR) adoption to better care is mixed. More nuanced measures of adoption, particularly those that capture the common incremental approach of adding functions over time in US hospitals, could help elucidate the relationship between adoption and outcomes. We used data for the period 2008–13 to assess the relationship between EHR adoption and thirty-day mortality rates. We found that baseline adoption was associated with a 0.11-percentage-point higher rate per function. Over time, maturation of the baseline functions was associated with a 0.09-percentage-point reduction in mortality rate per year per function. Each new function adopted in the study period was associated with a 0.21-percentage-point reduction in mortality rate per year per function. We observed effect modification based on size and teaching status, with small and nonteaching hospitals realizing greater gains. These findings suggest that national investment in hospital EHRs should yield improvements in mortality rates, but achieving them will take time.

Lin et al. (2018). EHRs and Patient Mortality. Health Affairs.

Mixed Evidence with Respect to EHRs on Hard Patient Outcomes?

Research and Applications

Electronic health records systems and hospital clinical performance: a study of nationwide hospital data

Neal Yuan,[1]* R. Adams Dudley,[2] W. John Boscardin,[3] and Grace A. Lin[2,4]

[1]Cedars-Sinai Smidt Heart Institute, Los Angeles, California, USA, [2]Center for Healthcare Value, Philip R. Lee Institute for Health Policy Studies, University of California, San Francisco, California, USA, [3]Department of Epidemiology and Biostatistics and Division of Geriatrics, University of California at San Francisco and San Francisco Veterans Affairs Medical Center, San Francisco, California, USA, and [4]Division of General Internal Medicine, University of California, San Francisco, California, USA

*Corresponding Author: Neal Yuan, MD, Cedars-Sinai Smidt Heart Institute, 8700 Beverly Blvd, Davis 1015, Los Angeles, CA 90048, USA (Neal.Yuan@cshs.org)

Received 8 January 2019; Revised 11 May 2019; Editorial Decision 14 May 2019; Accepted 18 May 2019

ABSTRACT

Objective: Electronic health records (EHRs) were expected to yield numerous benefits. However, early studies found mixed evidence of this. We sought to determine whether widespread adoption of modern EHRs in the US has improved clinical care.

Methods: We studied hospitals reporting performance measures from 2008–2015 in the Centers for Medicare and Medicaid Services Hospital Compare database that also reported having an EHR in the American Hospital Association 2015 IT supplement. Using interrupted time-series analysis, we examined the association of EHR implementation, EHR vendor, and Meaningful Use status with 11 process measures and 30-day hospital readmission and mortality rates for heart failure, pneumonia, and acute myocardial infarction.

Results: A total of 1246 hospitals contributed 8222 hospital-years. Compared to hospitals without EHRs, hospitals with EHRs had significant improvements over time on 5 of 11 process measures. There were no substantial differences in readmission or mortality rates. Hospitals with CPSI EHR systems performed worse on several process and outcome measures. Otherwise, we found no substantial improvements in process measures or condition-specific outcomes by duration of EHR use, EHR vendor, or a hospital's Meaningful Use Stage 1 or Stage 2 status.

Conclusion: In this national study of hospitals with modern EHRs, EHR use was associated with better process of care measure performance but did not improve condition-specific readmission or mortality rates regardless of duration of EHR use, vendor choice, or Meaningful Use status. Further research is required to understand why EHRs have yet to improve standard outcome measures and how to better realize the potential benefits of EHR systems.

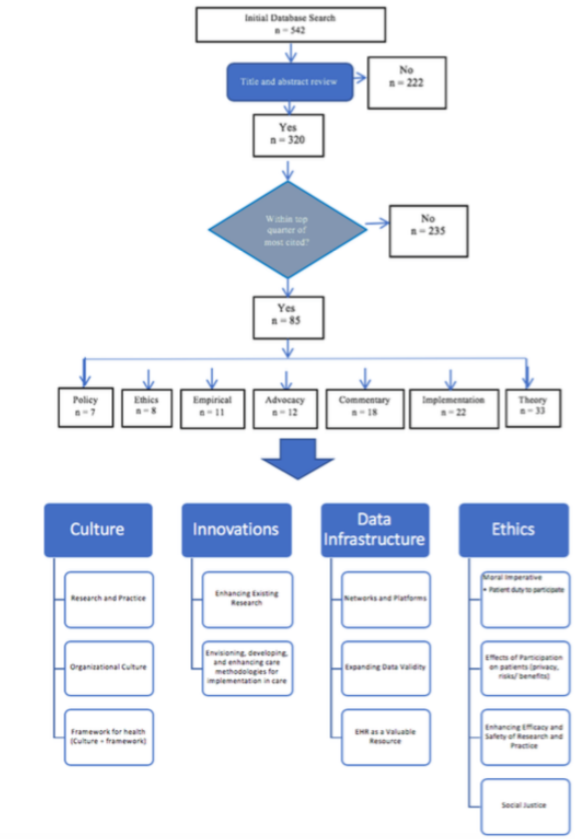
Key words: electronic health record, process measures, hospital readmission, hospital mortality, meaningful use

Yuan et al. (2019). EHRs and Hospital Performance. JAMIA.

EHRs as Essential Data Infrastructure for a Learning Health Care System

- Continued adoption and meaningful use of EHRs in clinical practice.
- Improved inter-operability EHR software across health system.
- Integration of EHRs with patient reported data, other RCD, registries, etc.
- Need to improve data quality. Design and methods fit for purpose.
- Embedded research/QI for continuous evidence, innovation, quality, value in healthcare.

Figure 2. Summary of search strategy and themes.



Platt et al. (2020). Learning Health System: A Decade Later. JMIR.

Part 2: EHRs for Secondary Data Analysis

EHRs for Secondary Data Analysis

SUMMARY POINTS LIST

1. EHRs provide researchers with low-cost sources of rich longitudinal health data on large geographically, socioeconomically, and culturally diverse populations.
2. EHRs offer particular advantages for environmental and social epidemiology where patient addresses can be linked to individual and contextual exposures that vary spatially.
3. The use of EHRs for epidemiology requires consideration of unique issues related to study population definition, population attrition, disease/case definition, and privacy concerns.
4. Major areas of population health EHR research include reevaluating prior findings; capitalizing on large sample sizes to analyze subgroups and to study rare diseases or multiple diseases simultaneously; social and environmental epidemiology; research on stigmatized conditions; predictive modeling; and exploiting natural experiments.
5. Future developments in EHRs including increased use and sophistication, improved capture of social and behavioral determinants of health, better standardization to allow data merging across health systems, and linkage to vital records and to other emerging technologies (e.g., personal sensing) and data streams (e.g., air pollution data, clinical biobanks) will improve data quality and expand research opportunities to improve public health.

Casey, et al (2016). EHRs for Public Health Research. Ann Review Public Health.

Developing EHR Research Questions

- Desc, Est, HypTest, Prediction, Causal.
- Target population. EHR vs. Data sample.
- What data is available? How measured?
- Sources of bias? Analytic challenges.

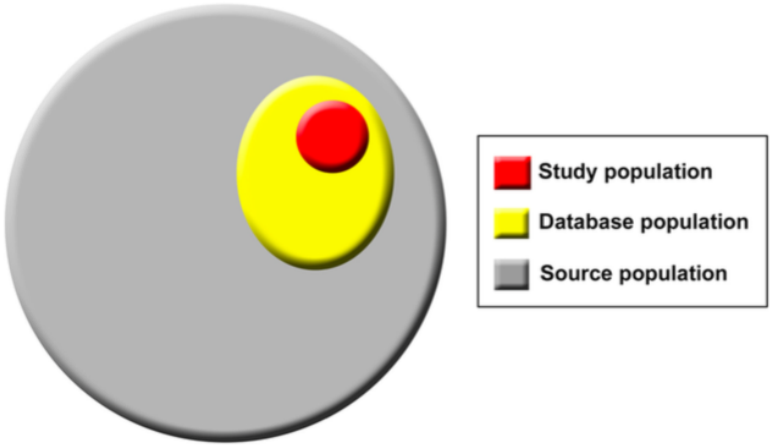
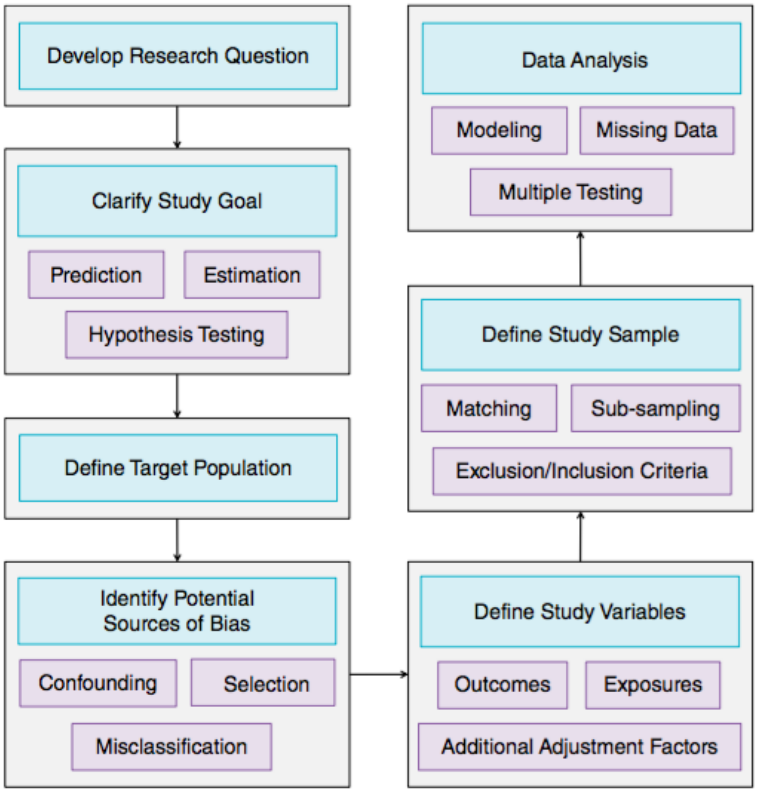


Fig 1. Population hierarchy in studies using routinely collected data sources.

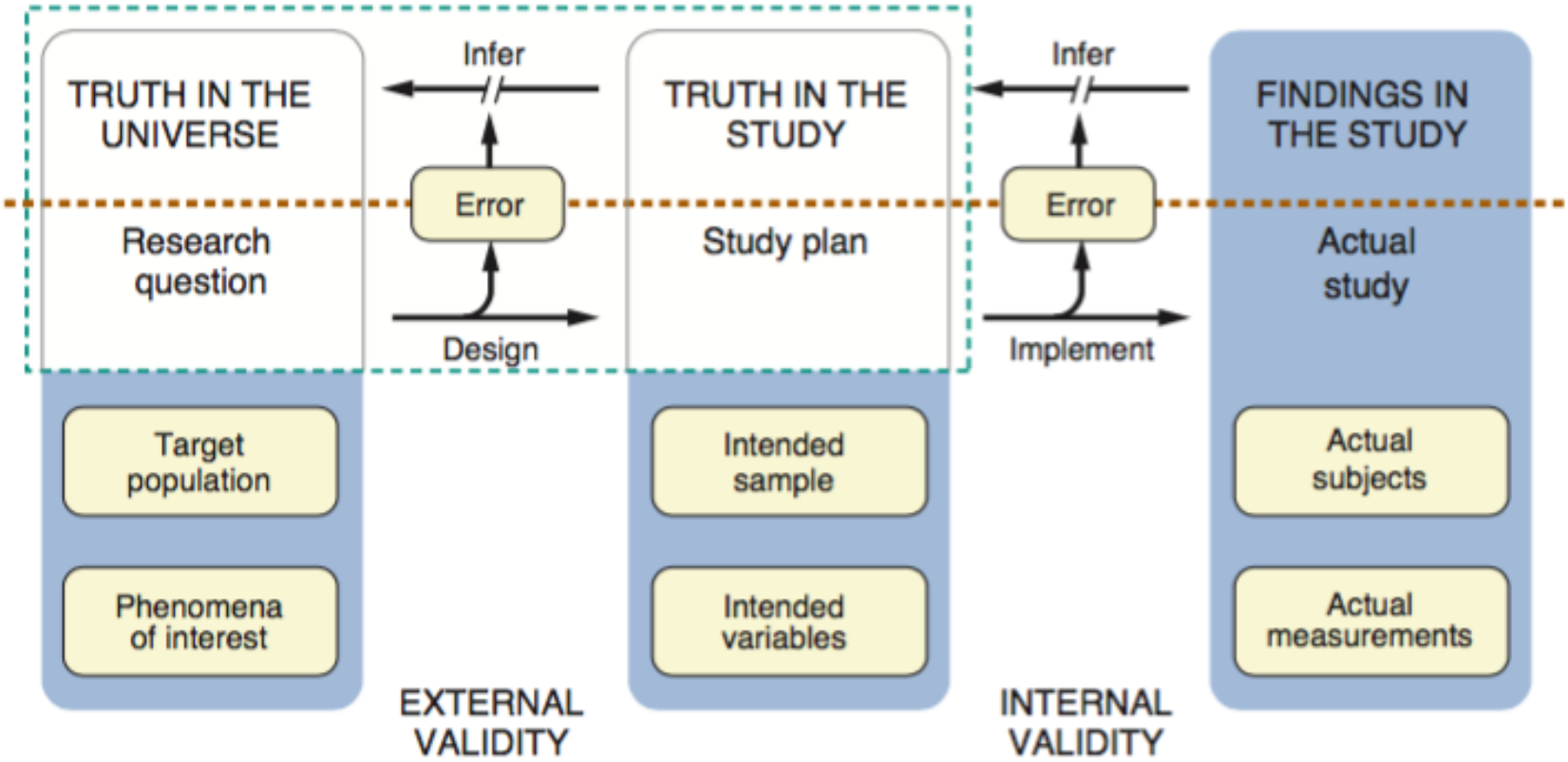
An Example EHR Research Pipeline



Beasley et al, (2018). Emerging Landscape BioBanks Linked to EHRs. Stats in Medicine.
Benchimol et al, (2020). RECORD Statement. PLoS One.

Clinical Research Frameworks - Applicable to EHRs

- Conceptualization of idealized vs. implemented study designs.
- Emphasize control over sources of internal/external validity.



■ **FIGURE 2.1** This chapter focuses on the area within the dashed green line, the challenge of choosing a research question that is of interest and can be tackled with a feasible study plan.

Hulley, et al (2018). Designing Clinical Research. Ch 2, p.12.

Design Issues and Other Potential Sources Bias in EHR Research

Box 1. Design Considerations for EHR Based Studies

- Where in the health system are the data collected?
- What is the coverage/catchment area of your health system?
- Is the patient population receiving care across multiple institutions/centers?
- Do the data constitute different catchments? (Admixture)
- How are you defining exposures and outcomes? (Phenotyping)
- How are you defining person-time?
  - What is an appropriate 'burn-in' period to define a cohort?
  - Is a 'burn-out' period necessary to define censoring?
- Do different populations produce more information (i.e. sicker patients have more encounters)?

Phelan et al. (2017). Illustrating Informed Presence Bias. eGEM.

Design Issues and Other Potential Sources Bias in EHR Research

Textbox 1. Possible sources of bias in the use or reuse of electronic health record data that have to be incorporated in the choice of research methods and interpretation of results.

1. Health care system bias, emanating from:
   - Reimbursement system, pay for performance parameters
   - Role of general practitioner in the health care system; gatekeeping/nongatekeeping
   - Professional clinical guidelines
   - Ease of access by patients to their records
   - Data sharing between health care providers
2. Practice workload
3. Variations between electronic health record (EHR) system functionalities and lay-out
4. Coding systems and thesauruses
5. Knowledge and education regarding the use of EHR systems
6. Data extraction tools
7. Data processing—redatabasing
8. Research dataset preparation
9. Research methodologies

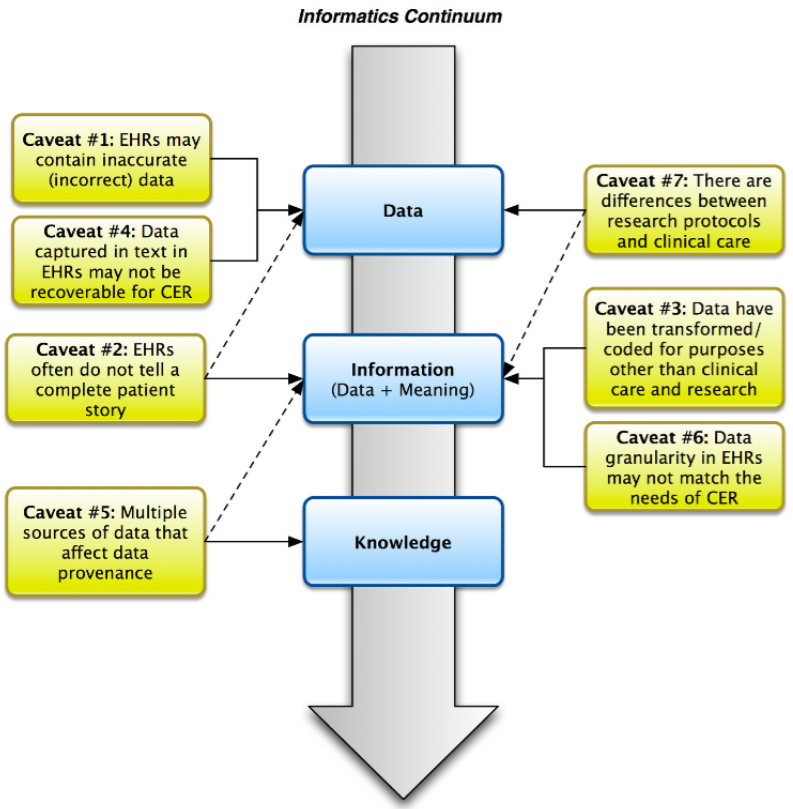
Verheij et al. (2018). Possible Sources of Bias in Primary Care EHRs. JMIR.

Design Issues and Other Potential Sources Bias in EHR Research

Table 1

Data idiosyncrasies for use of operational electronic health record data in comparative effectiveness research.

Type	Description	Example
Diagnostic uncertainty	- Diagnosis may be recorded when there is only a suspicion of disease	- Patient with suspected diabetes mellitus before diagnosis confirmed by laboratory testing
	- Some overlapping clinical conditions are difficult to distinguish reliably	- Various forms of upper respiratory and related infections, e.g., sinusitis, pharyngitis, bronchitis, rhinitis, etc.
	- Patients may only partially fit diagnostic criteria	- Patients with non-diagnostic gastrointestinal symptoms may partially fit diagnostic criteria for one or multiple diseases
	- Patients in whom diagnostic testing is done but is negative are still more likely to have disease	- Patients undergoing echocardiography for shortness of breath and edema who are found to have normal left ventricular function are different from asymptomatic patients with normal left ventricular function
Diagnostic timing	- Repeated diagnosis codes over time may represent a new event or a follow up to an prior event	- Two hospitalizations with a primary diagnosis of MI are likely two events, but a code for myocardial infarction in outpatient setting is more likely a follow up to an inpatient MI
	-First diagnosis in a database is not necessarily an incident case of disease	- A new patient in the system with diabetes may have had diabetes for many years prior to presentation
	- Chronic diseases may vary in severity over time	- Patient with congestive heart failure with waxing and waning of symptoms
Treatment choice and timing	- Many conditions do not require immediate drug or other treatment	- Hyperlipidemia or hypertension may have a trial of lifestyle changes before initiation of drug therapy
	- Patient co-morbidities may effect timing and choice of treatment	- Patient with hypertension may have related diagnoses that were not recorded before initiation of treatment, but may be recorded later to indicate the compelling reason for a treatment choice, such as the use of ACE inhibitors in hypertensive patients with heart failure
Treatment decisions	- Treatment decisions not randomized	- Physician choosing treatment based on personal views or biases regarding efficacy
	- Some treatment decisions are remote to the patient-provider interaction	- Restrictions by patient insurance or institutional drug formulary
	- Some treatments not reliably recorded	- Medications available over the counter and not requiring a prescription may not be recorded, e.g., aspirin, proton pump inhibitors
Treatment follow-up	- Some treatments confounded by clinical factors unrelated to condition being treated	- Patient with multiple co-morbidities may be seen more frequently and have conditions treated faster, e.g., hyperlipidemia in otherwise healthy person versus patient with diabetes and its complications
	- Non-clinical factors impact availability of data	- Patient access to resources in order to follow treatment recommendations may be limited due to travel, payor systems, or other non-clinical factors



Hersh et al. (2013). EHRs and Cost Effectiveness Research. Medical Care.

Design Issues and Other Potential Sources Bias in EHR Research

- Confounding (regression, stratification, matching). Causal estimands.
- Complex dependencies (hierarchical, temporal, spatial, crossed).
- Missing data: MCAR, MAR, NMAR.
- Multiple testing. Phishing. p-hacking.
- Selection Bias.
- Informative Visit Bias (Informative Visit Processes; Irregular Visit Processes).
- Misclassification Bias (Outcome vs. Covariate; Informative vs. Non-informative).

Little & Rubin. (2002). Statistical Analysis with Missing Data. Wiley.
Haneuse et al. (2016). Missing Data Mechanisms EHR Research. Epidemiology.
Farcomeni. (2008). Review Modern Multiple Hypothesis Testing. SMMR.
Rothman. (1990). No Adjustment Needed for Multiple Comparisons. Epidemiology.
Haneuse et al. (2016). Gen Framework Selection Bias EHRs. eGEMs.
Goldstein et al. (2019). Controlling Informed Presence Bias EHRs. AJE.
Pullenayagum et al. (2014). Longitudinal data irregular observation: rev methods. Stat Med.
Neuhaus et al. (2018). Outcome Dependent Visit Processes. Stat Med.

EHR Research Design

Underlying process (data generating mechanism) for EHR data sampling is dynamic/open cohort.

Common Sampling Processes in Observational EHR Research

- Cross-sectional. Descriptive/analytic.
- Prospective/retrospective cohort. Panel cohorts.
- Case-Control.
- Case Cohort.
- Nested Case Control.
- Two-stage sampling processes.
- Descriptive time series designs.

Ernster. (1994). Nested Case Control Studies. Preventative Medicine.
Breslow. (2010). Using Whole Case Cohort. AJE.
Barlow et al. (1999). Analysis of Case Cohort Designs. JCE.
Zeger et al. (2006). Time Series Analysis Public Health Data. Ann Rev Pub Health.
Diggle. (1999). Time Series and Biostatistics. Oxford Press.

EHR Research Design

Underlying process (data generating mechanism) for EHR data sampling is dynamic/open cohort.

Experimental Designs in EHR Research

- Individual randomized RCTs (no so common).
- (Pragmatic) cluster RCTs.
- Stepped wedge designs (and other cross-over variants).
- Platform adaptive designs.
- Zelen designs and informed consent.

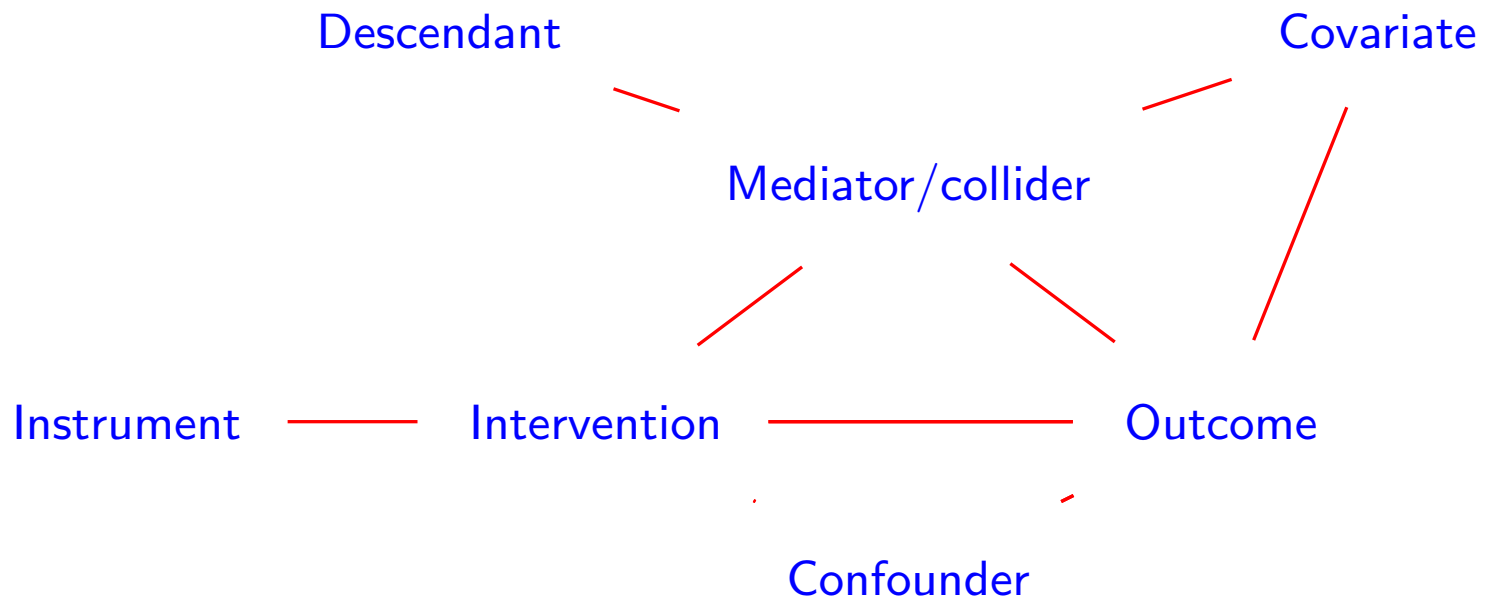
Natural Experiments in EHR Research

- Time series designs. Intervention analysis.
- Regression discontinuity designs.

Hussey et al. (2007). Design Analysis Stepped Wedge Trials. Comtemp Clin Trials.
Zwarenstein et al. (2009). Making trials matter: pragmatic and explanatory trials. Trials.
Butler et al. (2018). Response-adaptive platform trials in primary care. BJGP.
Torgerson et al. (1998). What is Zelens Design. BMJ.
Wagner et al. (2002). Segmented regression analysis of ITS studies. J Clin Ther Res.
Pankratz. (1991). Forecasting with Dynamic Reg Models. Wiley.
Campbell & Stanley. (1963). Experimental and Quasi-Experimental Designs for Research.

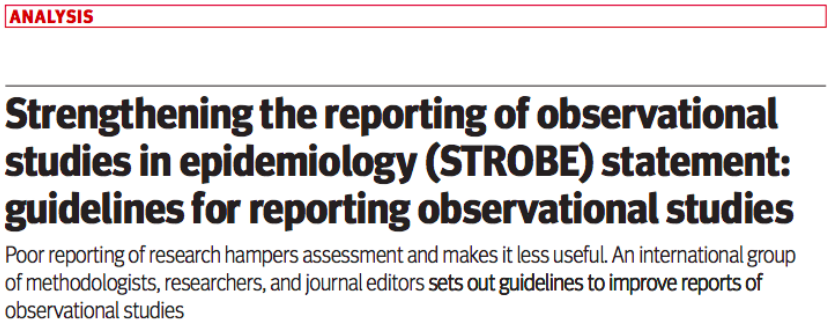
Causal Inference from Observational EHR Designs

- Propensity scores. Instrumetal variable analysis. Mediation analysis.
- Causal Factor Analysis and SEM. Causal DAGs and NPSEM.



Rosenbaum. (2002). Observational Designs. Springer.
Hernan & Robins. (2019). Causal Inference: What If.
Bollen. (1989). Confirmatory Factor Analysis. Wiley.
Pearl. (2000). Causality.

Reporting Guidelines for EHR Studies: STROBE and RECORD



ANALYSIS

Strengthening the reporting of observational studies in epidemiology (STROBE) statement: guidelines for reporting observational studies

Poor reporting of research hampers assessment and makes it less useful. An international group of methodologists, researchers, and journal editors sets out guidelines to improve reports of observational studies

GUIDELINES AND GUIDANCE

The REporting of studies Conducted using Observational Routinely-collected health Data (RECORD) Statement

Eric I. Benchimol[1,2]*, Liam Smeeth[3], Astrid Guttmann[2,4], Katie Harron[3], David Moher[5], Irene Petersen[6], Henrik T. Sørensen[7], Erik von Elm[8‡], Sinéad M. Langan[3‡]*, RECORD Working Committee[¶]

1 Children's Hospital of Eastern Ontario Research Institute, Department of Pediatrics and School of Epidemiology, Public Health and Preventive Medicine, University of Ottawa, Ottawa, Canada, **2** Institute for Clinical Evaluative Sciences, Toronto, Canada, **3** London School of Hygiene and Tropical Medicine, London, United Kingdom, **4** Hospital for Sick Children, Department of Paediatrics and Institute of Health Policy, Management and Evaluation, University of Toronto, Toronto, Canada, **5** Ottawa Hospital Research Institute, Ottawa, Canada, and School of Epidemiology, Public Health and Preventive Medicine, University of Ottawa, Ottawa, Canada, **6** Department of Primary Care and Population Health, University College London, London, United Kingdom, **7** Department of Clinical Epidemiology, Aarhus University, Aarhus, Denmark, **8** Cochrane Switzerland, Institute of Social and Preventive Medicine, University of Lausanne, Lausanne, Switzerland

‡ The senior authors contributed equally to this work.
¶ Membership of the RECORD Working Committee is provided in the Acknowledgments.
* ebenchimol@cheo.on.ca (EIB); sinead.langan@lshtm.ac.uk (SML)

Abstract

Routinely collected health data, obtained for administrative and clinical purposes without specific a priori research goals, are increasingly used for research. The rapid evolution and availability of these data have revealed issues not addressed by existing reporting guidelines, such as Strengthening the Reporting of Observational Studies in Epidemiology (STROBE). The REporting of studies Conducted using Observational Routinely collected health Data (RECORD) statement was created to fill these gaps. RECORD was created as an extension to the STROBE statement to address reporting items specific to observational studies using routinely collected health data. RECORD consists of a checklist of 13 items related to the title, abstract, introduction, methods, results, and discussion section of articles, and other information required for inclusion in such research reports. This document contains the checklist and explanatory and elaboration information to enhance the use of the checklist. Examples of good reporting for each RECORD checklist item are also included herein. This document, as well as the accompanying website and message board (http://www.record-statement.org), will enhance the implementation and understanding of RECORD. Through implementation of RECORD, authors, journals editors, and peer reviewers can encourage transparency of research reporting.

---

von Elm et al, (2007). STROBE Statement. BMJ.
Benchimol et al, (2020). RECORD Statement. PLoS One.

Selected Topics in EHR Informatics

Selected Topic 1

Designing & Evaluating EHR Phenotyping Studies

Phenotyping Electronic Health Records

- Phenotype specifies observable, possibly time-varying state of a patient.
- Phenotypes derived from EHR can be outcomes, covariates, incl/excl criterion, etc.
- Synonyms: case finding algorithms, code set engineering, etc.

Table 1

Applications of electronic phenotyping across study types

Study type	Use cases
Cross-sectional	Epidemiological research
	Hospital administration/resource allocation
	Adherence to diagnostic/treatment guidelines
	Quality measurement
Association (case-control/cohort)	Genome-wide association studies
	Pharmacovigilance
	Identifying clinical risk factors and protective factors
	Clinical decision support
	Clinical effectiveness research
	Predictive modeling
Experimental	Clinical trial recruitment
	Pragmatic trials
	Adaptive/randomized, embedded, multifactorial, adaptive platform trials

Banda et al. (2018). Phenotpying: Rule Based vs. ML. Ann Rev Biomedical Data Sci.

Phenotyping Design/Evaluation

- Rule/algo identify case positive/negative patients in EHR.
- How to know if phenotype algorithm any good?

Cross-sectional Sampling

- (Large) random sample EHR.
- Expert annotated gold-standard labels.
- Development/evaluation of phenotype algorithm.
- Limitations: cost/time, rare conditions, etc.

Case-Control (Two-stage) Sampling

- Identify most case positive patients in EHR.
- Random sample case negative patients EHR.
- Development/evaluation of phenotype algorithm.
- Limitations: cost/time, re-calibration, etc.

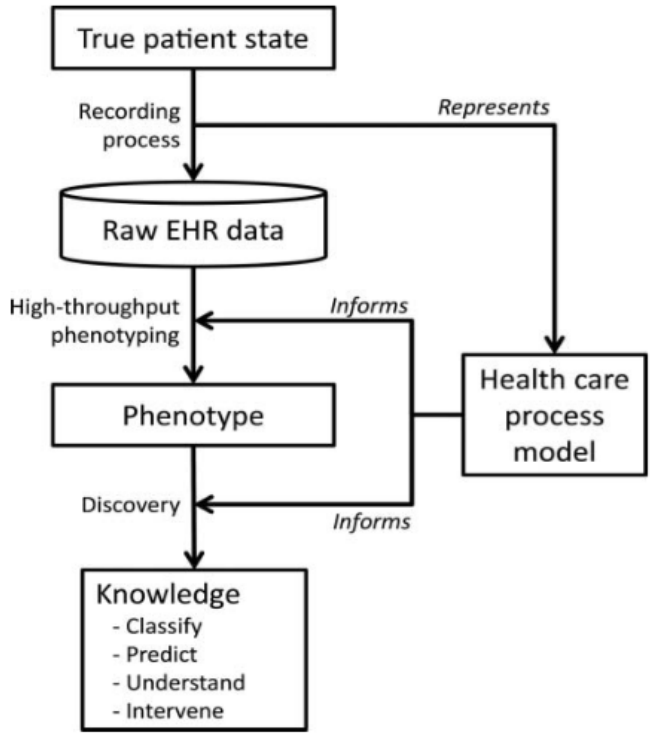


Figure 2 Phenotyping and discovery. The raw electronic health record (EHR) data are an indirect reflection of the true patient state due to the recording process. Attempts to create phenotypes and discover knowledge must account for the recording. The healthcare process model represents the salient features of the recording process and informs the phenotyping and discovery.

Hripcsak et al. (2012). Next Gen Phenotyping EHRs. JAMIA.

Annotation

- Two or more experts review and label charts for algorithm development.
- Major rate limiting step: timely and costly.
- Assess inter/intra-rater agreement (kappa, etc.).

Algorithm Development

- Rule based methods. Boolean logical queries. Expert derived, consensus timely.
- Supervised ML. Feature engineering timely. Overfit small train samples.
- Increasingly novel designs: attempt circumvent annotation/development/overfitting costs.
  - Unsupervised learning.
  - Semi-supervised learning.
  - Learning with Noisy Labels (noise tolerant learning).
  - Active learning.

Evaluating Phenotype Algorithms

- Held out test set evaluation?
- Metrics: sensitivity, specificity, PPV, NPV, F1, etc.
- Generalizability? Geographic, temporal, etc.
- Portability, transparency, etc.

Figure 1. General model of automatic phenotyping algorithms.

Alzoubi et al. (2019). Auto Phenotyping EHRs: A Review. MDPI Electronics.

Trends in EHR Phenotyping Studies

- Rule based methods popular. Interpretable. Fast. Lack awareness language/context?
- Supervised ML gains popularity. Flexible, context aware. Black-box?
- Neural network language models (attention, transformers, BERT, etc.) state of art?
- Non-traditional designs considered to reduce time and cost of development/evaluation.
- Newer methods incorporate external knowledge bases and ontologies.
- Emphasis towards shareable phenotypes algorithms and/or modular/transportable pipelines.
  - PheKB, eMERGE, OHDSI, CALIBRE, etc.

Review Methods (Banda et al)

Primary method	Number of papers
Rule-based	19
Natural language processing	35
Standard machine learning	25
Learning with noisy data	11
Unsupervised phenotype discovery	11
Hybrid approaches	3
Collaborative frameworks	10
Total	89

Review Data Structures (Alzoubi et al)

Table 2. Categorise of feature extraction techniques.

Feature Extraction Method	Papers
Structured only	[18,35,38,90–96]
BoW only	[30,49]
Keyword search only	[29,54,97]
Concept extraction only	[31,59,69,72,73,75,98]
Structured + BoW	[42,50,52,99]
Structured + Keyword search	[6,16,17,32,55,58,60]
Structured + Concept extraction	[15,28,36,37,48,68,70,71,74,100]

Hripcsak et al. (2012). Next Gen Phenotyping EHRs. JAMIA.
Shivade et al. (2014). Review Identify Phenotype Cohorts EHRs. JAMIA.
Banda et al. (2018). Phenotpying: Rule Based vs. ML. Ann Rev Biomedical Data Sci.
Alzoubi et al. (2019). Auto Phenotyping EHRs: A Review. MDPI Electronics.

Novel Designs: Active Learning (Chen et al, 2014)

- Goal: learn high quality phenotype algorithm with smaller amount labelled data.
- How: (oracle) presents statistical algorithm with "most useful" features for learning.
- Caveat: what is the approach for identifying optimal learning instances from data set?

Active Learning Pipeline

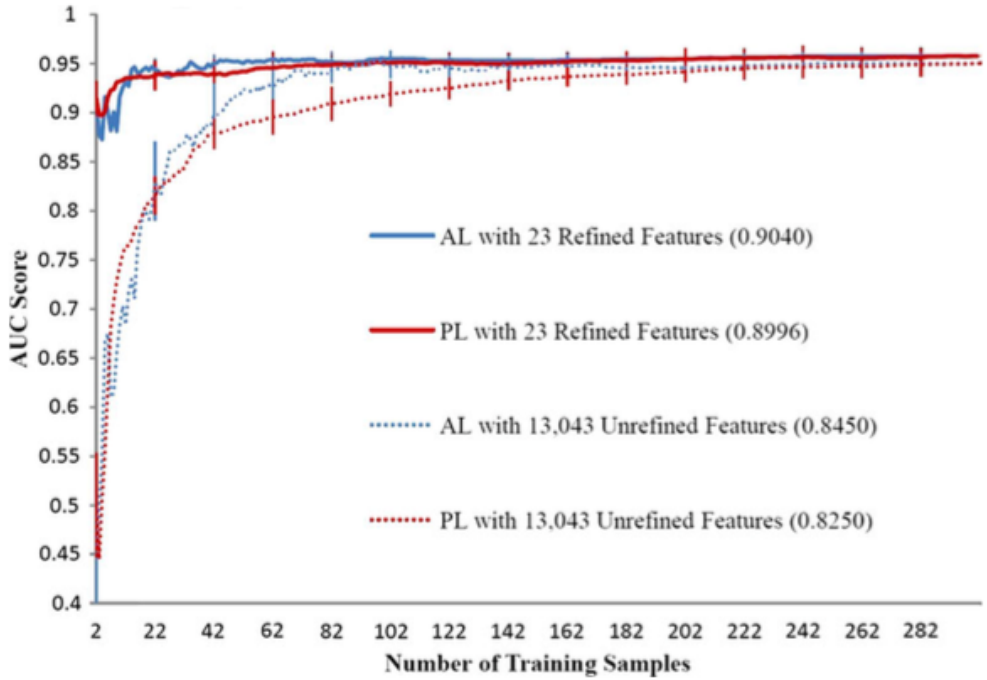
AL for ML-based phenotyping

When building ML classifiers for phenotyping, the conventional method is to select a set of samples randomly for annotation, which is also called passive learning (PL). On the contrary, AL uses a querying algorithm actively to select the most informative samples for annotation in an iterative and interactive fashion. In this study, we simulated a pool-based AL strategy[25] in ML-based phenotyping tasks using existing datasets. We assumed that all samples of a phenotype (eg, 376 samples for RA) were unlabeled and they formed the pool for AL. We then took the following steps:

1. Initialize—we randomly selected two samples from the pool (one control and one case) and used them to build the initial ML model.
2. Predict—we used the trained ML model to predict remaining samples in the pool.
3. Query—we used the uncertainty sampling algorithm[30] to find most informative samples, which are samples that have the most uncertain predicted labels by the ML model. For binary SVM classifiers, uncertainty sampling-based algorithm queries the samples that are closest to the hyperplane, which separates cases and controls. We queried one sample in each iteration.
4. Annotate and re-train the ML model—we annotated the next most informative sample from step 3 (by assigning labels from gold standard) and re-built the ML model by combining the new training sample with previous ones.
5. Repeat and stop—we repeated steps 2–4 until the stop criterion was met. In this study, the learning process stopped when the unlabeled pool was empty.

For PL, we ran the same procedure using random sampling as its querying algorithm.

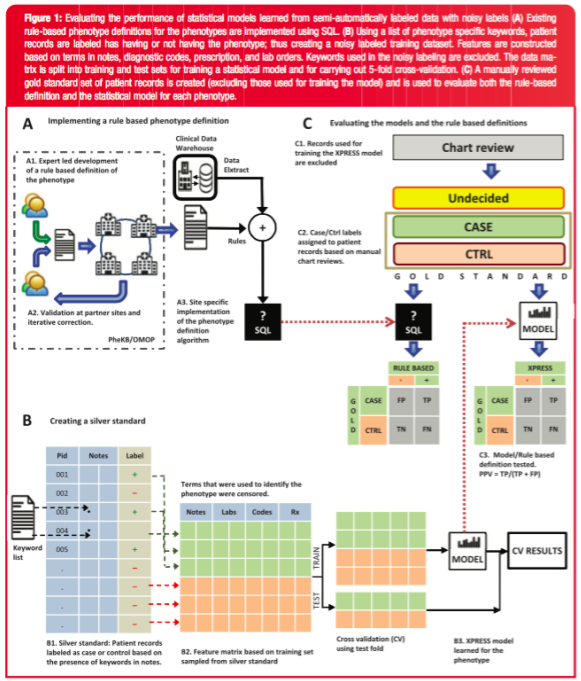
Comparison Active/Traditional Learning



Chen et al. (2014). Active Learning High Throughput Phenotyping EHRs. JAMIA.

Novel Designs: Learning with Noisy Labels

- Goal: learn high quality phenotype algorithm with smaller amount labelled data.
- How: noisy positive/negative labels for training, based on anchor method (silver standard).
- Anchor is specific code/phrase/concept with high PPV (possibly low sensitivity).
- Amount noisy labels needed proportional to error/noise rate silver standard ($\frac{1}{(1-2\tau)^2}$).



Agarwal et al. (2016). Learn Stat Models Phenotyping from Noisy Labelled Data. JAMIA.
Halpern et al. (2016). EHR Phenotyping using Anchor and Learn Method. JAMIA.
Banda et al. (2018). EHR Phenotyping with APHRODITE in OHDSI. JAMIA.

Novel Designs: Learning with Noisy Labels & Feature Engineering

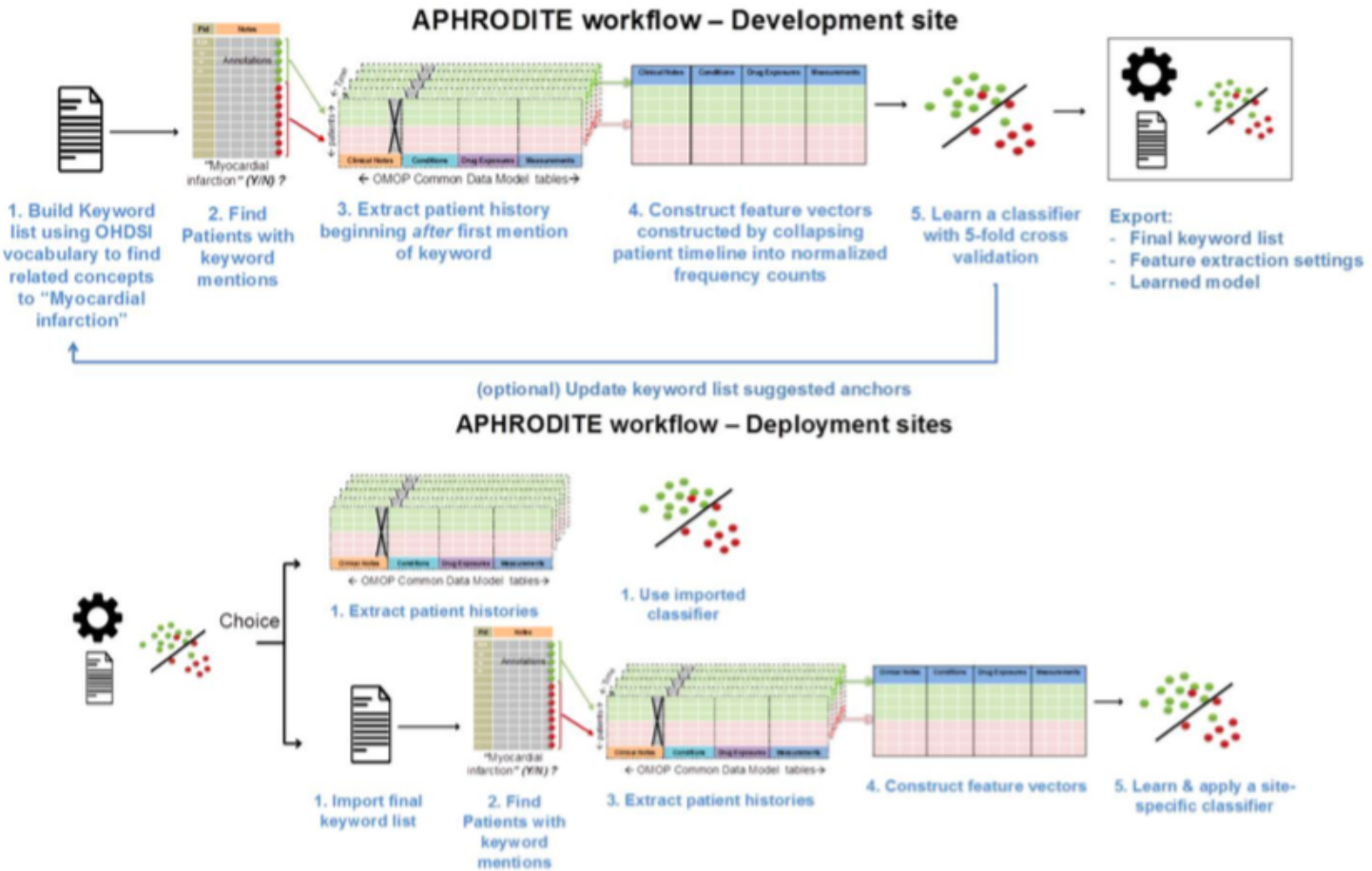


Figure 1. APHRODITE phenotype development/deployment framework schematics. Phenotype definitions are initially learned at development sites and exported for deployment. At deployment sites, users have a choice to use the final keyword list to learn their own site-specific models or use the pre-built classifier.

Banda et al. (2018). EHR Phenotyping with APHRODITE in OHDSI. JAMIA.

Novel Designs: Learning with Noisy Labels & Feature Engineering

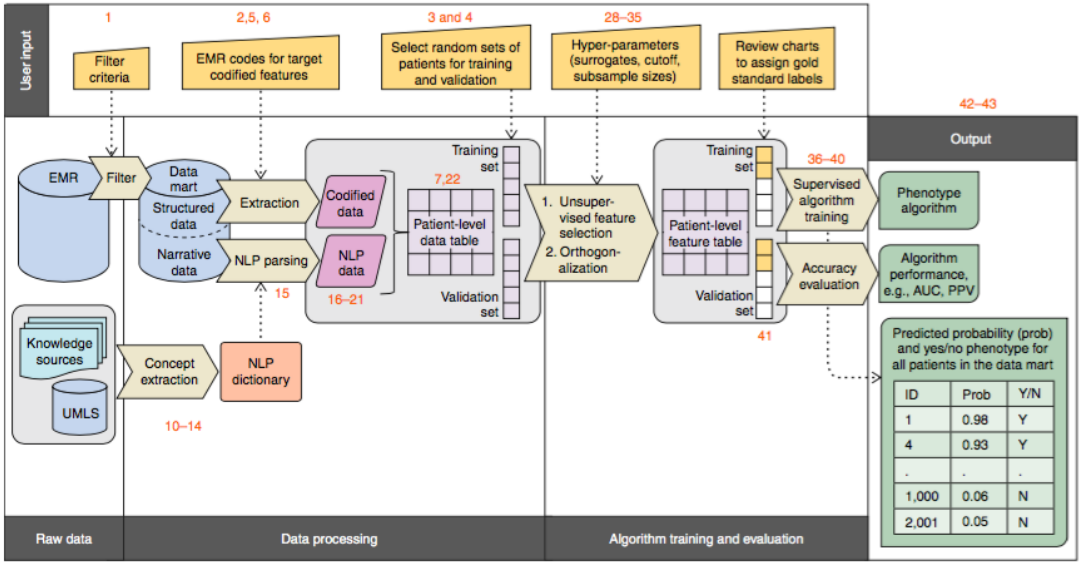


Fig. 4 | Detailed flow of PheCAP protocol. User inputs required at various steps in the PheCAP protocol are specified at the top of the figure as the protocol moves from data extraction, data processing, and algorithm training/validation to the final outputs: a phenotype algorithm, a probability of the phenotype for all subjects in the data mart, and a classification of the phenotype for each subject (yes or no). Orange numbers in the figure correspond to Procedure steps.

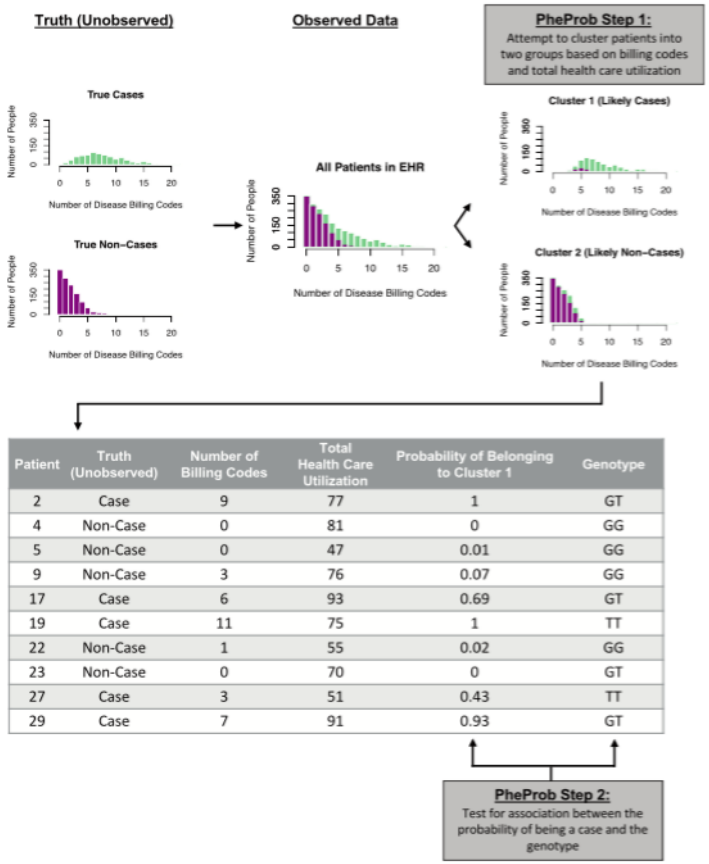
Timing

Steps 1 and 2, creation of the EMR data mart: ~24 h (variable depending on data size and infrastructure)
Steps 3 and 4, chart review and acquisition of gold standard labels: 1 week, depending on the availability of domain expert
Steps 5–7, identification and extraction of codified data features from the structured EMR data for the algorithm: ~12 h
Steps 8 and 9, preparation of analysis environment in R: <1 h
Steps 10–14, concept collection for candidate NLP features: ~12 h
Steps 15–22, note parsing to obtain NLP feature data: ~20+ h; can vary widely depending on size of dataset and size of NLP dictionary
Steps 23–27, loading of EMR data into R: <1 h
Steps 28–35, performing SAFE: <1 h
Steps 36–40, supervised algorithm training: <1 h
Step 41, algorithm validation: <1 h
Steps 42 and 43, application of the model: <1 h

Zhang et al. (2019). PheCAP: Phenotyping Common Semi-Sup Approach. Nature Protocols.

Novel Designs: Unsupervised Phenotyping

- Goal: learn high quality phenotype algorithm with no amount labelled data.
- How: surrogate/anchor features, and unsupervised learning algorithms.



Patient	Truth (Unobserved)	Number of Billing Codes	Total Health Care Utilization	Probability of Belonging to Cluster 1	Genotype
2	Case	9	77	1	GT
4	Non-Case	0	81	0	GG
5	Non-Case	0	47	0.01	GG
9	Non-Case	3	76	0.07	GG
17	Case	6	93	0.69	GT
19	Case	11	75	1	TT
22	Non-Case	1	55	0.02	GG
23	Non-Case	0	70	0	GT
27	Case	3	51	0.43	TT
29	Case	7	91	0.93	GT

Figure 1. Workflow of the PheProb method. True disease status of patients is unknown; instead, the number of billing codes for each disease is observed. The PheProb method clusters individuals based on billing codes, and tests for an association between a genetic marker such as a SNP and the clustering-based probability of being a case.

Sinnott et al. (2018). EHR Phenotyping using Anchor and Learn Method. JAMIA.
Yu et al. (2017). Enabling Phenotypic Big Data with PheNORM. JAMIA.

Novel Designs: Unsupervised Phenotyping

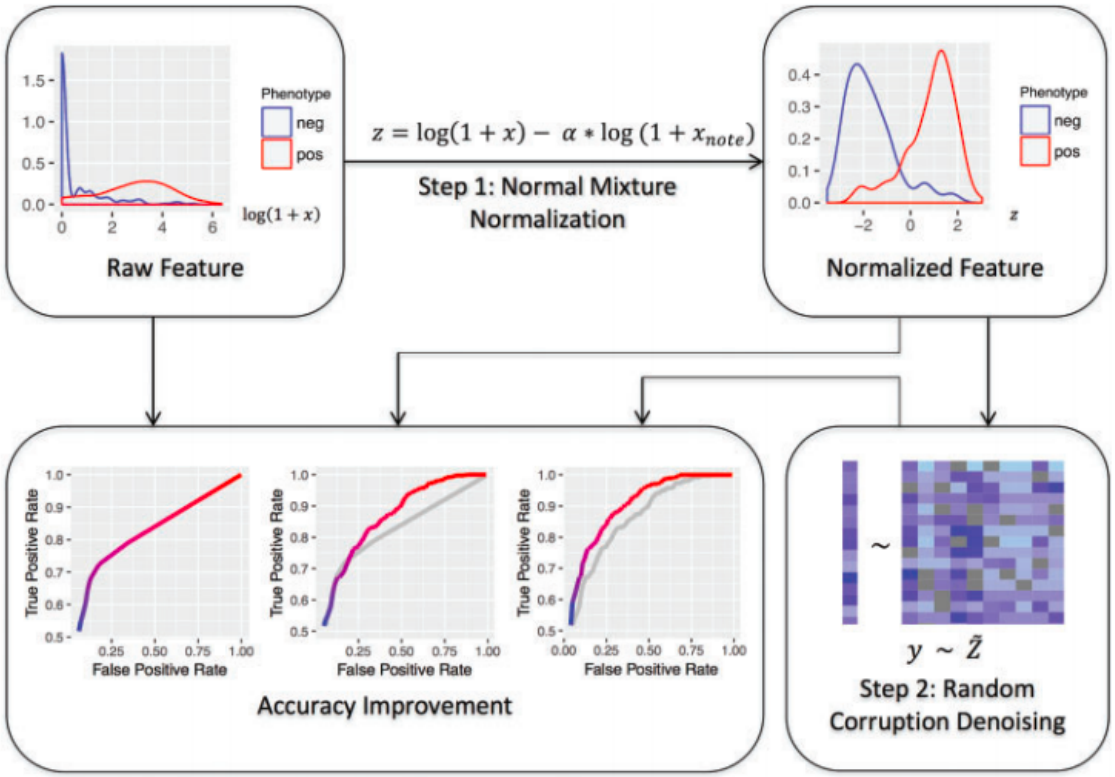
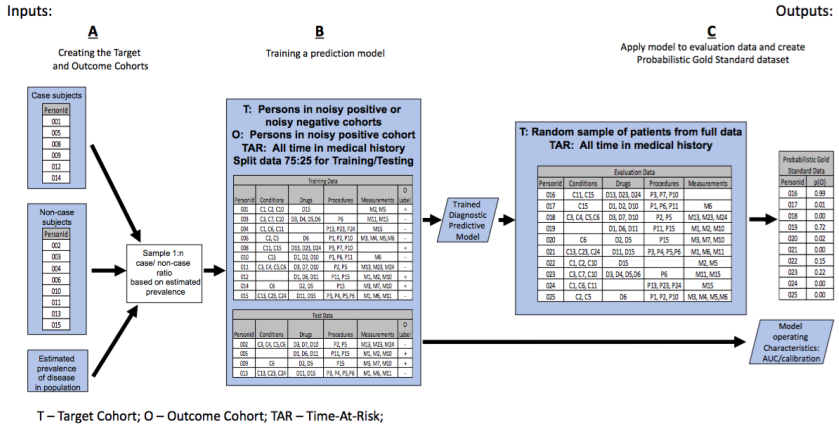


Figure 1. Workflow of PheNorm. Top left: density plot (after logarithm transformation) of a highly predictive feature (illustrated here using the ICD-9-CM count of ulcerative colitis from a Partners HealthCare EHR datamart), denoted by x, in patients who do (the right curve) and do not (the left curve) have the phenotype. Top right: Density plot of the ICD-9-CM count after the normal mixture transformation using the total number of notes in the patient's EHR, denoted by x_{note}. The densities of the phenotype positive and negative patients are approximately normally distributed, and the 2 populations are separated to a large degree. Bottom right: The transformed feature is denoised by self-regression of the transformed feature, denoted by y, onto the entire transformed and randomly corrupted feature set, denoted by $\tilde{Z}$ with dropout. The transformed features are then combined into a prediction formula for disease status classification based on the estimated regression coefficient. Bottom left: The receiver operating characteristic (ROC) curve of the feature or score in each step, with AUC growing steadily (gray curves are copies of the ROC curves from the previous steps).

Yu et al. (2017). Enabling Phenotypic Big Data with PheNORM. JAMIA.

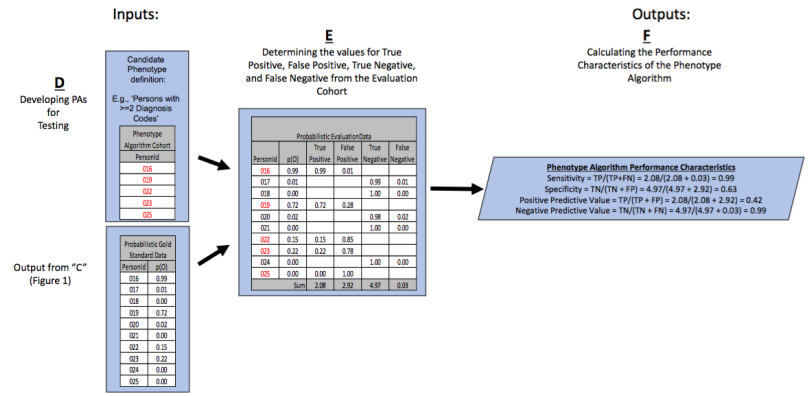
Novel Designs: Learning with Noisy Labels & Zero Label Evaluation

- Goal: learn high quality phenotype algorithm with no amount labelled data.
- How: surrogate/anchor features for noisy model learning.
- How: Extend definition evaluation metrics, allowing zero labels.
- Caveat: Are evaluation metrics these designs valid/unbiased?



T – Target Cohort; O – Outcome Cohort; TAR – Time-At-Risk;

Fig. 1. The first portion of the PheValuator process flow creating the target and outcome cohorts to be used in training the diagnostic predictive model for developing the probabilistic gold standard phenotype data set to be used in phenotype algorithm evaluation.



p(O) – Probability of Outcome; TP – True Positive; FN – False Negative; TN – True Negative; FP – False Positive

Fig. 2. The last portion of the PheValuator process flow using test phenotype algorithms along with the probabilistic gold standard phenotype data for developing the performance characteristics of the phenotype algorithm.

Swerdel et al. (2019). PheValuator: Devt and Evaluation Phenotyping Algorithms. JBI.

How do we Know if a Proposed Phenotyping Algorithm is Any Good?

- Good design. Strong evaluation. Transparent reporting.
- All phenotype algorithms should be evaluated to be trusted!! See RECORD statement.
- Caution: Many phenotyping strategies. Not all created equal. Susceptibility bias?

Phenotype Reporting (Chen et al)

Recommendations for Research	Recommendations for Reporting
Determine whether cohort definitions are precise enough to establish a reference standard, such as through assessment by multiple reviewers and their interrater reliability.	Describe the EHR data elements used to ascertain the presence or absence of a given variable (e.g., presence of an opioid drug prescription: any prescription record for a drug with the Anatomical Therapeutic Chemical category N02A, for opioid analgesics).
Evaluate the performance of algorithms for cohort identification (i.e., electronic phenotyping) to determine how many subjects an algorithm misses and how many are likely to be incorrectly included in a resulting data set.	Report the data elements, as well as the necessary and sufficient criteria to include or exclude a study subject.
This is consistent with RECORD guideline recommendations: 6.1: The methods of study population selection (such as codes or algorithms used to identify subjects) should be listed in detail. If this is not possible, an explanation should be provided. 6.2: Any validation studies of the codes or algorithms used to select the population should be referenced. If validation was conducted for this study and not published elsewhere, detailed methods and results should be provided.	If an executable algorithmic method was used to identify cohorts, define what reference standard was used and the performance metrics of human and algorithmic identification methods (e.g., precision, recall, and interrater kappa). This is consistent with RECORD guideline recommendations: 12.1: Authors should describe the extent to which the investigators had access to the database population used to create the study population. 12.2: Authors should provide information on the data cleaning methods used in the study.

Biases in Evaluation? (Hong

We assume the validation sample is a random sample of all patients. However, in practice, subjects may be selected into the validation sample due to convenience sampling or stratified sampling, which may result in biased estimates of accuracy if the estimation methods do not properly account for non-randomness in the ascertainment. This bias is known as work-up bias (Ransohoff and Feinstein, 1978) and verification bias (Begg and Greenes, 1983). Future work is needed to account for the verification bias in the validation set.

Chen et al. (2020). Reporting Obs Studies Using EHR Data. Ann Int Med.
Benchimol et al. (2020). RECORD Statement. PLoS One.
Hong et al, (2017). Semi-Sup Validation EHR Phenotyping. Biometrics.

Selected Topic 2

Characterizing Primary Care Using Clinical Text Data and Topic Models

Problem: Vast Amounts of Clinical Text Data...What is it About?

Matrix factorization/decomposition of input DTM (or TCM):

- Tokenize input clinical text. Represent text as "bag of words".
- Document Term Matrix: rows are document specific term-frequency vector.
- Document Term Matrix: D documents. V words in vocabulary.
- LSA/NMF: K dimensional embedding for document/vocabulary axes.
- LSA/NMF: k=1...K rows of ϕ represent topical/archetypical vectors.
- LSA/NMF: d=1...D rows of θ represent patient affinity to topics.

$$\underset{D \times V \text{matrix}}{\begin{bmatrix} w_{1,1} & \cdots & \cdots & \cdots & w_{1,V} \\ \vdots & & & & \vdots \\ w_{D,1} & \cdots & \cdots & \cdots & w_{D,V} \end{bmatrix}} = \underset{D \times K \text{matrix}}{\begin{bmatrix} \theta_{1,1} & \cdots & \theta_{1,K} \\ \vdots & & \vdots \\ \theta_{D,1} & \cdots & \theta_{D,K} \end{bmatrix}} * \underset{K \times V \text{matrix}}{\begin{bmatrix} \phi_{1,1} & \cdots & \cdots & \cdots & \phi_{1,V} \\ \vdots & & & & \vdots \\ \phi_{K,1} & \cdots & \cdots & \cdots & \phi_{K,V} \end{bmatrix}}$$

Deerwester et al. (1990). Latent Semantic Analysis. JASIS.

Lee et al, (1999). NMF: Finding Parts in Objects. Nature.

Topic Models:

- Describe thematic structure of corpora via latent topics.
- A topics is a discrete/categorical distribution over vocabulary.

Mixed Membership Models (MMMs): Mixture Models for Grouped Data

- Data naturally grouped ($w_{d,n}$). For $d = 1...D$ and $n = 1...N_d$.
- Each group represented as mixture model.
- Mixture components shared across groups.
- Mixture proportions vary between groups.

Generative Process for Mixed Membership Models

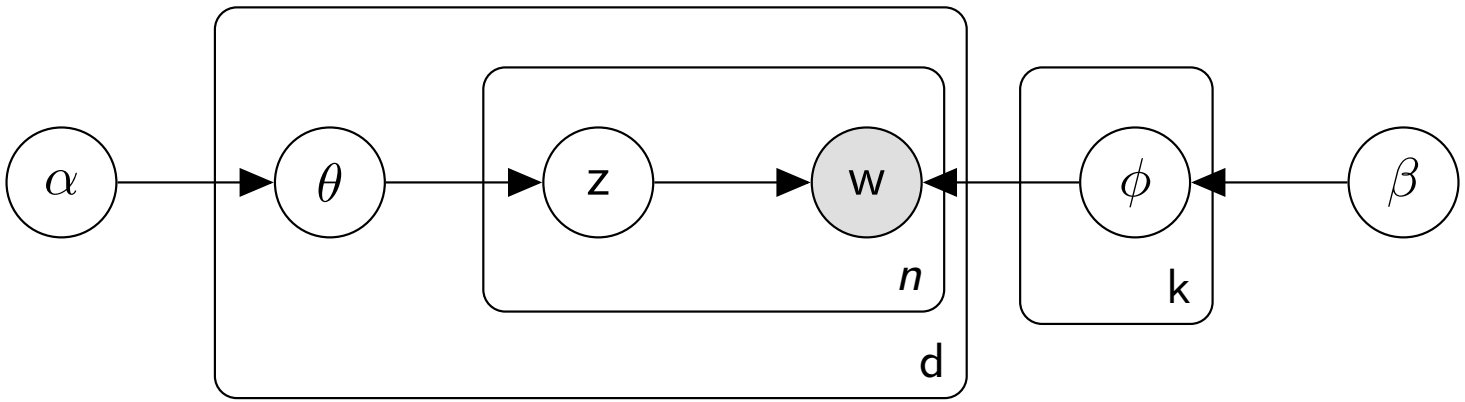
1. Draw shared components $\phi_k \sim f(\cdot|\beta)$
2. For each group $d = 1...D$:
   1. Draw proportions $\theta_d \sim \text{Dir}(\alpha)$
   2. For each data point, $w_{d,n}$, $n = 1...N_d$:
      1. Draw a mixture assignment $z_{d,n} \sim \text{Cat}(\theta_i)$
      2. Draw the data point $w_{d,n} \sim g(\cdot|\phi_{z_{d,n}})$

Blei et al. (2003). Latent Dirichlet Allocation. JMLR.
Griffiths et al, (2004). Finding Scientific Topics. Science.

Latent Dirichlet Allocation (LDA): a MMM for discrete/text data.

1. Draw shared components $\phi_k \sim \text{Dir}(\cdot|\beta)$
2. For each group $d = 1 \ldots D$:
   1. Draw proportions $\theta_d \sim \text{Dir}(\alpha)$
   2. For each data point, $w_{d,n}$, $n = 1 \ldots N_d$:
      1. Draw a mixture assignment $z_{d,n} \sim \text{Cat}(\theta_i)$
      2. Draw the data point $w_{d,n} \sim \text{Cat}(\cdot|\phi_{z_{d,n}})$

Latent Dirichlet Allocation (LDA): a graphical model perspective.



Blei et al. (2003). Latent Dirichlet Allocation. JMLR.

Latent Dirichlet Allocation (LDA): Joint/posterior Distribution

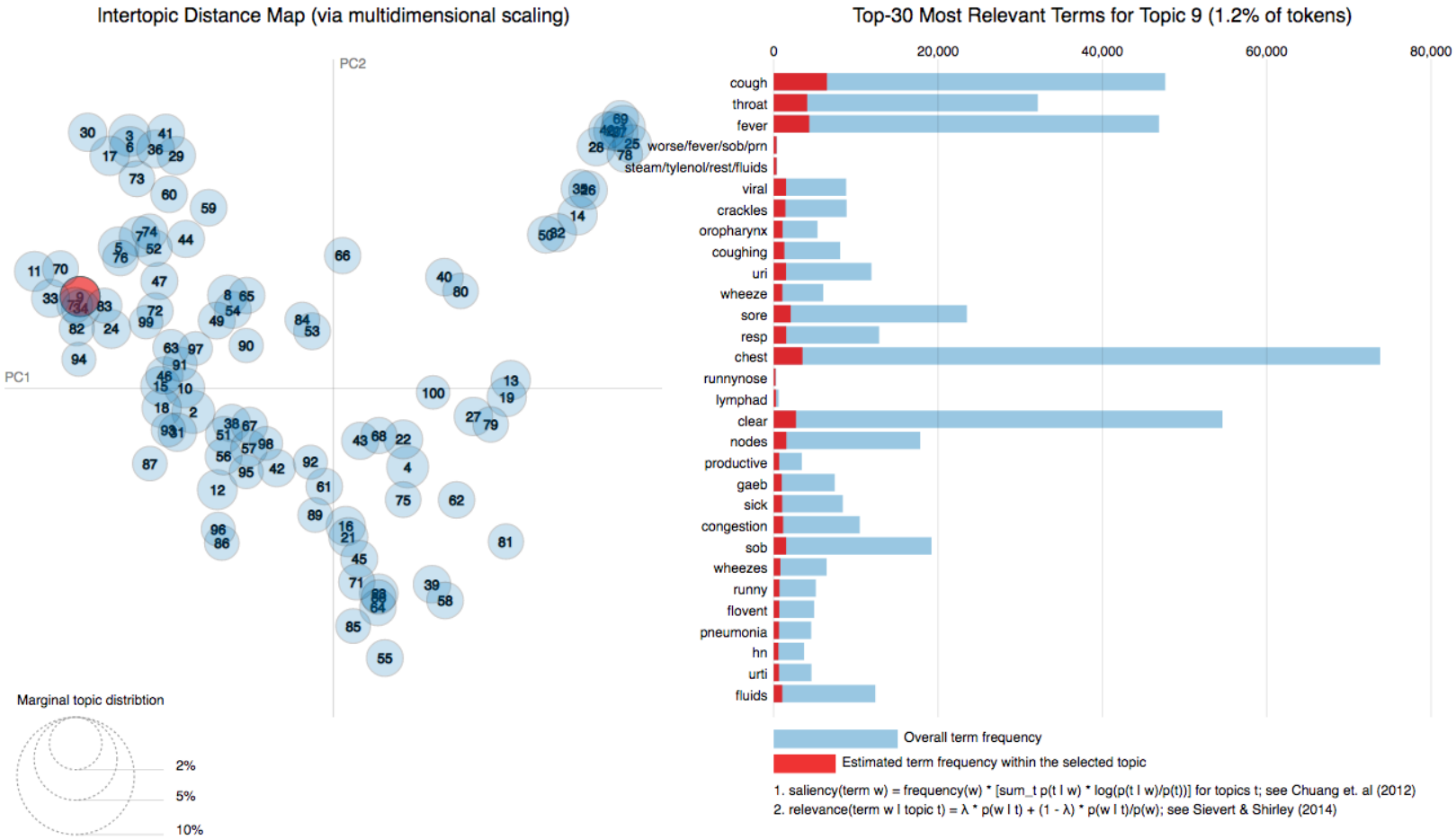
- Generative-process/graphical-model imply joint distribution
- Posterior distribution from joint distribution, by conditional probability.

$$
\begin{aligned}
P(z, \theta, \phi | w, \alpha, \beta) &= \frac{P(w, z, \theta, \phi | \alpha, \beta)}{P(w)} \\
&\propto P(w, z, \theta, \phi | \alpha, \beta) \\
&= \prod_k P(\phi_k; \beta) \prod_d P(\theta_d; \alpha) \prod_n P(z_{d,n} | \theta_d) P(w_{d,n} | \phi_{z_{d,n}}) \\
&= \prod_k \text{Dir}(\phi_k; \beta) \prod_d \text{Dir}(\theta_d; \alpha) \prod_n \text{Cat}(z_{d,n} | \theta_d) \text{Cat}(w_{d,n} | \phi_{z_{d,n}})
\end{aligned}
$$

Latent Dirichlet Allocation: Posterior Inference

- Gibbs Sampling (Pritchard et al, 2000).
- Collapsed Gibbs Sampling (Griffiths, et al, 2002, 2004, 2007).
- Variational Inference (Blei et al, 2003, 2016; Hoffman et al, 2013)

Demo of LDAvis Topic Viewer App on NYFHT Progress Note Corpus



Potential Applications of Topic Models in Clincial and Public Health

- Surveillance (temporal, spatial).
- Unsupervised phenotyping.
- Text as outcomes in designed experiments.
- Improved organization, retrieval, browsing in clinical EHR applications.

Selected Topic 3

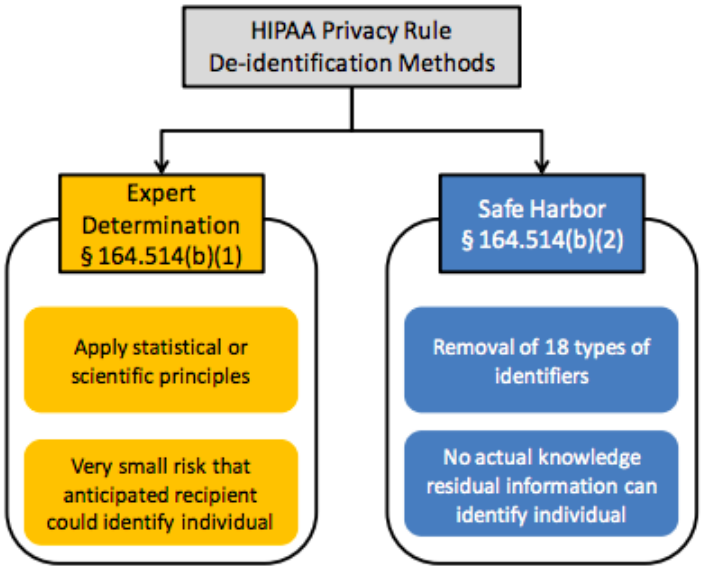
De-identification of Protected Health Information from Clinical Text Data

Rationale for De-identification

> The increasing adoption of health information technologies in the United States accelerates their potential to facilitate beneficial studies that combine large, complex data sets from multiple sources. The process of de-identification, by which identifiers are removed from the health information, mitigates privacy risks to individuals and thereby supports the secondary use of data for comparative effectiveness studies, policy assessment, life sciences research, and other endeavors.
>
> — Health Insurance and Portability and Protection Act (1996)

Goals for De-identification Systems

- Reduce re-identification risk, replacing PHI with realistic surrogates.
- Maintain medically relevant information and biomedical semantic structure from text data.
- Use privacy/security as a tool to enhance public trust in the secondary use of medical data.



Who Defines PHI and What to De-Identify?

- HIPAA (USA) specified two approaches to DEID: (1) expert determination, (2) safe-harbor.
- Other jurisdictions lack explicit PHI definitions (PHIPAA-Canada; GDPR-EU).
- Do we need more risk averse interpretations of HIPAA? Is HIPAA enough?

No.	PHI Type
1	Names
2	All geographic subdivisions smaller than a state
3	Dates
4	Telephone Numbers
5	Vehicle Identifiers
6	Fax Numbers
7	Device Identifiers and Serial Numbers
8	Emails
9	URLs
10	Social Security Numbers
11	Medical Record Numbers
12	IP Addresses
13	Biometric Identifiers
14	Health Plan Beneficiary Numbers
15	Full-face photographic images and any comparable images
16	Account Numbers
17	Certificate/license numbers
18	Any other unique identifying number, characteristic, or code.

Approaches to De-Identification of Medical Notes

- Complete redaction of PHI.
- Substitution with granular identification tokens.
- Substitution with realistic surrogate information.

Replacements for Example Note			
Original	Substituted	Identifier	Redacted
Dear Janine Keane, as we have discussed, I hereby send you the requested information about my patient, Julie Andrews. You can reach her via email (her address is jandrews@gmail.com) or via phone: 998 785 6756. Sincerely, Elijah Hunt, MD	Dear Rosie Copeland, as we have discussed, I hereby send you the requested information about my patient, Beatrice Burton. You can reach her via email (her address is lebewukak@halava.fi) or via phone: (836) 230-3149. Sincerely, Jayden Bush, MD	Dear <name>, as we have discussed, I hereby send you the requested information about my patient, <name>. You can reach her via email (her address is <email>) or via phone: <phone>. Sincerely, <name>, MD	Dear **********, as we have discussed, I hereby send you the requested information about my patient, **********. You can reach her via email (her address is **********) or via phone: **********. Sincerely, **********, MD

Bruckhardt et al. (2017). deidentify. AMIA.

Methods for de-identification of PHI

- Rule based systems.
  - REGEX, pattern matching, dictionaries.
- Machine learning systems.
  - Conditional random fields popular.
  - Require heavy feature engineering.
- Hybrid ML/rule-based systems.
- Deep learning systems.
  - Little/no feature engineering required.
  - RNNs, bi-LSTMs, Transformers.

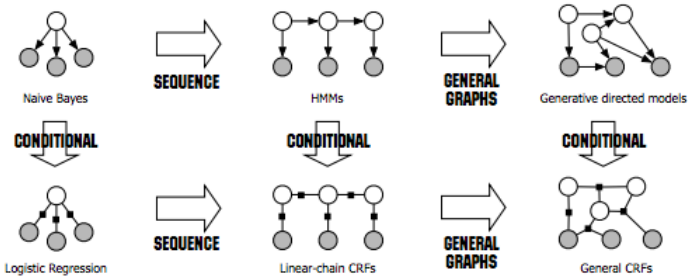


Figure 1.2 Diagram of the relationship between naive Bayes, logistic regression, HMMs, linear-chain CRFs, generative models, and general CRFs.

Examples of Engineered Features

Bag-of-words: unigrams, bigrams and trigrams of words within a window of [-2, 2].
Part-of-speech (POS) tags: unigrams, bigrams and trigrams of POS tags within a window of $[-2, 2]$. The Stanford POS Tagger [42] was used for POS tagging.
Combinations of words and POS tags: combining current word with the unigrams, bigrams and trigrams of POS tags within a window of $[-1, 1]$, i.e. w_0p_{-1}, w_0p_0, w_0p_1, $w_0p_{-1}p_0$, $w_0p_0p_1$, $w_0p_{-1}p_1$, $w_0p_{-1}p_0p_1$, where w_0, p_{-1}, p_0 and p_1 denote current word, last, current and next POS tags respectively.
Sentence information: number of words in current sentence, whether there is an end mark at the end of current sentence such as '.', '?' and '!', whether there is any bracket unmatched in current sentence.
Affixes: prefixes and suffixes of length from 1 to 5.
Orthographical features: whether the word is upper case, contains uppercase characters, contains punctuation marks, contains digits, etc.
Word shapes: mapping any or consecutive uppercase character (s), lowercase character(s), digit(s) and other character(s) in current word to ' A', ' a', ' #' and ' -' respectively. For instance, the word shapes of "Hospital" are "Aaaaaaaa" and "Aa".
Section information: twenty-nine section headers (see the supplementary file) were collected manually such as "History of Present Illness"; we check which section current word belongs to.
General NER information: the Stanford Named Entity Recognizer [43] was used to generate the NER tags of current word, include: person, date, organization, location, and number tags, etc.
Word representation features: two types of word representation features generated by Brown clustering [44] and word2vec [45] on training sets and a large unlabeled MEDLINE corpus.
Dictionary features: four categories of localization dictionaries: COUNTRY, STATE, CITY and ZIP were collected from Internet, and each word was labeled with ' 0' or ' 1' by dictionary lookup.

Liu et al. (2017). DEID with RNNs and CRFs. JBI.
Sutton et al. (2011). Intro Conditional Random Fields. Foundations Trends ML.

Benchmark Data for DEID Research

- Uzuner (2007). Eval SOTA Auto-DEID.
  - Discharge summaries, Parters HC.
  - 889pt, 500k tokens (20k PHI).
- Neamatullah (2008). Auto-DEID EMRs.
  - Sample nursing notes MIMIC-II.
  - 169pt, 2k note. 330k tok (2k PHI).
- Stubbs (2014). Long Clin Narr DEID.
  - Longitudinal notes diab/cvd pats.
  - 300pt, 1k nt. 800k tok (30k PHI).
- Stubbs (2016). DEID Psych Adm Notes.
  - Psychiatric intake notes Partners.
  - 1k records, 1.8M tok (35k PHI).
- Uzuner (2017). RNNs for DEID.
  - MIMIC-III ICU discharge note.
  - 1600pt, 3M tokens (61k PHI).

Table 2
Comparison of token counts in 2016 and 2014 shared task corpora.

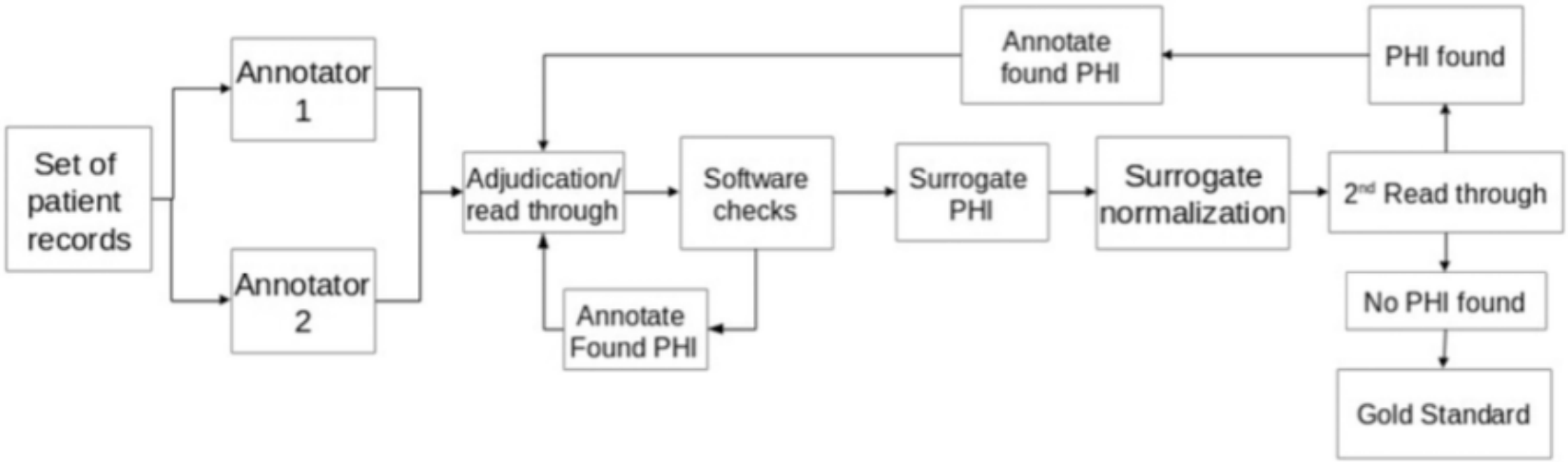
	2016	2014
Total tokens	1,862,452	805,118
Average per record	1,862.4	617.4
Max	4610	2984
Min	304	617

Table 3
PHI category distributions between 2016 training and testing data, and comparison of PHI category totals between 2016 and 2014 corpora.

PHI category	Total #: 2016	Total #: 2014
NAME: PATIENT	2107	2195
NAME: DOCTOR	3963	4797
NAME: USERNAME	25	356
PROFESSION	2481	413
LOCATION: HOSPITAL	3523	2312
LOCATION: ORGANIZATION	1810	206
LOCATION: STREET	80	352
LOCATION: CITY	2214	654
LOCATION: STATE	1143	504
LOCATION: COUNTRY	1042	183
LOCATION: ZIP CODE	40	352
LOCATION: OTHER	44	17
AGE	5991	1997
DATE	9544	12,487
CONTACT: PHONE	256	524
CONTACT: FAX	9	10
CONTACT: EMAIL	7	5
CONTACT: URL	8	2
CONTACT: IPADDRESS	0	0
ID: SSN	0	0
ID: MEDICAL RECORD	6	1033
ID: HEALTH PLAN	2	1
ID: ACCOUNT	0	0
ID: LICENSE	59	0
ID: VEHICLE	0	0
ID: DEVICE	0	15
ID: BIO ID	0	1
ID: ID NUMBER	10	456
Total # of PHI phrases	**34,364**	**28,872**
Average PHI per file	**34**	**22.14**

Uzuner et al. (2007). Eval State Art DEID. JAMIA.
Neahmatullah et al. (2008). Auto DEID Free Text Med Records. BMC Med Informatics.
Stubbs et al. (2014). DEID Narratives i2b2 2014. JBI.
Stubbs et al. (2016). DEID Psych Notes: CEGS-NGRID. JBI.
Dernocourt et al. (2017). DEID with RNNs. JAMIA.

The DEID Annotation Pipeline from i2b2 2014 (Stubbs et al, 2014).



Design Considerations for DEID Studies

- How many patients and notes to sample? Generalizability of sample?
- What definition of PHI, and how to define labels in annotation process.
- Evaluating the annotation process: inter/intra-rater agreement, time/cost, etc.
- Choice of algorithmic method for DEID: rule, ML, hybrid, DL, etc.
- Evaluation metrics for DEID: instance/token-level, BIO, SENS/PPV/F1.
- Train/val/test vs. Cross-val; Temporal/Geographic Hold-Outs; Generalizabily.
- Post-hoc error analysis. Are errors poor training data vs. poor model?
- Other issues: pre-annotation, PHI enrichment, over-scrubbing, etc.

Stubbs et al. (2014). DEID Narratives i2b2 2014. JBI.

Evaluation Metrics for DEID

- Sensitivity (Recall): TP/(TP+FN)
- Positive Predictive Value (Precision): TP/(TP+FP)
- F1 (harmonic mean recall/precision: (2*P*R)/(P+R)

Aggregation of Evaluation Metrics

- Micro-averaged. Estimate evaluation metrics for entire corpus.
- Macro-averaged. Estimate evaluation metrics for each document, and average.

Unit of Analysis

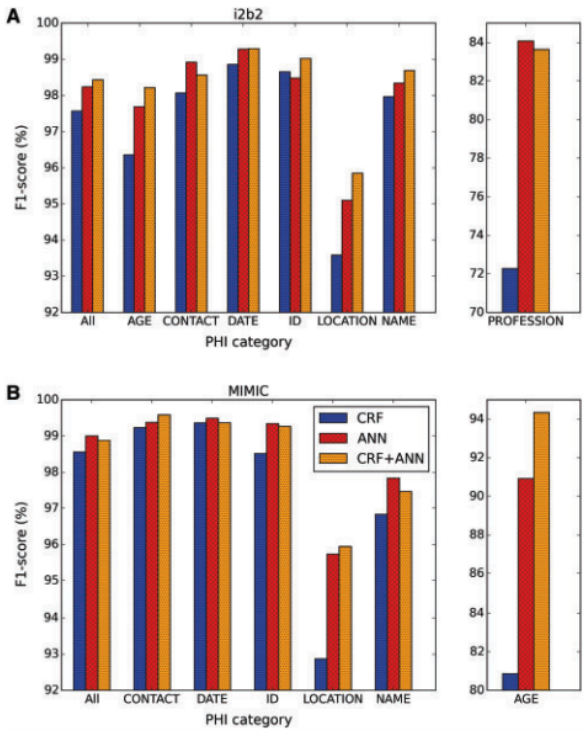
- Token level.
- Instance level.

Instance Level Evaluation:

- 3 error types: (1) substitution, (2) insertion, and (3) deletion.
- Instance-level PHI has 3 slots: (1) type, (2), content and (3) extent.
- 4 values per slot: (1) correct, (2) incorrect, (3) missing, and (4) spurious.
- Let C,I,D,S represent number correct, insertion/deletion/substitution errors.
- Instance level precision: C/(C+S+1)
- Instance level recall: C/(C+S+D)
- Instance level F1: (2*P*R)/(P+R)

Error Analysis for DEID Systems

- Errors by PHI type.
- Location/organization/profession hard.
- Name mid-easy.
- PHI with simple patterns easy.



To gain a better understanding of the best performing model and an intuition for its limitations, we conduct a manual error analysis of the false positives (FPs) and false negatives (FNs) produced by the BiLSTM-CRF on the test set. We discuss the error categorization scheme in Section 5.2.1 and present the results in Section 5.2.2.

5.2.1 Error Categorization. We distinguish between two error groups (1) modeling errors, and (2) annotation/preprocessing errors. We define modeling errors to be problems that can be addressed with different de-identification techniques and additional training data. In contrast, annotation and preprocessing errors are not directly caused by the sequence labeling model, but are issues in the training data or the preprocessing pipeline which need to be addressed manually. Inspired by the classification scheme of Dernoncourt et al. [5], we consider the following sources of modeling errors:

- **Abbreviation.** PHI instances which are abbreviations or acronyms for names, care institutes and companies. These are hard to detect and can be ambiguous as they are easily confused with medical terms and measurements.
- **Ambiguity.** A human reader may be unable to decide whether a given text fragment is PHI.
- **Debatable.** It can be argued that the token should not have been annotated as PHI.
- **Prefix.** Names of internal locations, organizations and companies are often prefixed with articles (i.e., "de" and "het"). Sometimes, it is unclear whether the prefix is part of the official name or part of the sentence construction. This ambiguity is reflected in the training data which causes the model to inconsistently include or exclude those prefixes.
- **Common Language.** PHI instances consisting of common language are hard to discriminate from the surrounding text.
- **Other.** Remaining modeling errors that do not fall into the categories mentioned above. In those cases, it is not immediately apparent why the misclassification occurs.

Preprocessing errors are categorized as follows:

- **Missing Annotation.** The text fragment is PHI, but was missed during the annotation phase.
- **Annotation Error.** The annotator assigned an invalid entity boundary.
- **Tokenization Error.** The annotated text span could not be split into a compatible token span. Those tokens were marked as "Outside (O)" during BIO tagging.

Trienes et al. (2018). Comparing Rule Based and Deep NN DEID. Arxiv.
Dernocourt et al. (2017). DEID with RNNs. JAMIA.

Systematic Review (Meystre, 2010): State of the Art is a Hybrid ML System.

Table 4 Algorithms and features used by systems mostly based on machine learning methods

De-identification system	Machine learning algorithm	Features		
		Lexical/morphological	Syntactic	Semantic
Aramaki	CRF	Word, surrounding words (5 words window), capitalization, word length, regular expressions (date, phone), sentence position and length.	POS (word + 2 surrounding words)	Dictionary terms (names, locations)
Gardner	CRF	Word lemma, capitalization, numbers, prefixes/suffixes, 2-3 character n-grams	POS (word)	None
Guo	SVM	Word, capitalization, prefixes/suffixes, word length, numbers, regular expressions (date, ID, phone, age)	POS (word)	Entities extracted by ANNIE (doctors, hospitals, locations)
Hara	SVM	Word, lemma, capitalization, regular expressions (phone, date, ID)	POS (word)	Section headings
Szarvas	Decision Tree	Word length, capitalization, numbers, regular expressions (age, date, ID, phone), token frequency	None	Dictionary terms (first names, US locations, countries, cities, diseases, non-PHI terms), section heading.
Taira	Maximum Entropy	Capitalization, punctuation, numbers, regular expressions (prefixes, physician and hospital name, syndrome/disease/procedure)	POS (word)	Semantic lexicon, dictionary terms (proper names, prefixes, drugs, devices), semantic selectional restrictions
Uzuner	SVM	Word, lexical bigrams, capitalization, punctuation, numbers, word length.	POS (word + 2 surrounding words), syntactic bigrams (link grammar)	MeSH ID, dictionary terms (names, US and world locations, hospital names), section headers.
Wellner	CRF	Word unigrams/bigrams, surrounding words (3 words window), prefixes/suffixes, capitalization, numbers, regular expressions (phone, ID, zip, date, locations/hospitals)	None	Dictionary terms (US states, months, general English terms).

Meystre et al. (2010). Automatic DEID: Review of Literature. BMC Medical Informatics.

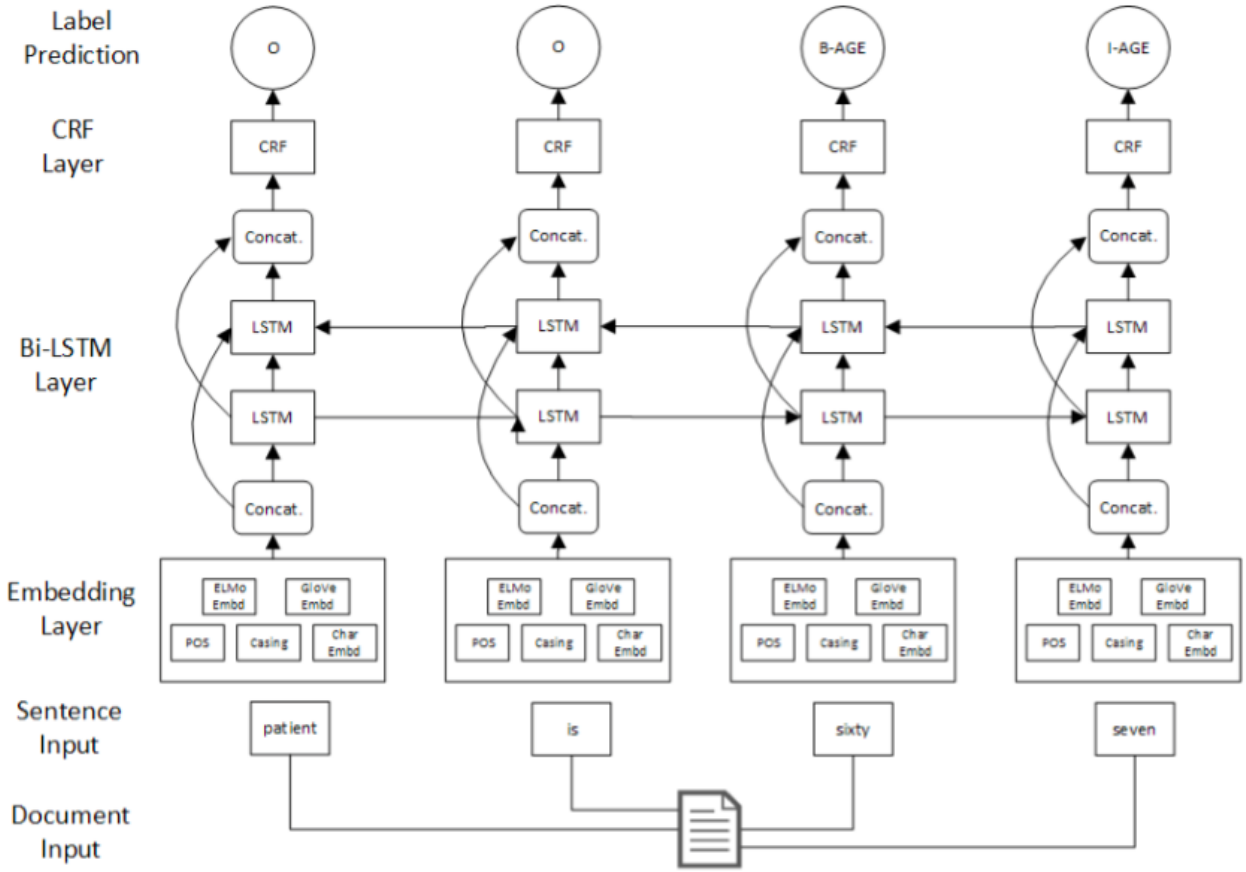
CEGS-NGRID (Stubbs, 2016): State of the Art is a Deep Learning Model

Rank	Team	Systems and processing	Precision	Recall	F1
1	Harbin Institute of Technology Shenzhen Graduate School	**Pre-processing**: Tokenization **System**: 4 modules: – CRF – 2 BI-LSTM – Rules **Post-processing**: merged CRF and BI-LSTM outputs with SVN ensemble classifier, then incorporated rules [20]	0.9422	0.8881	0.91430
2	University of Texas Health Science Center at Houston	**Pre-processing**: tokenization, POS tagging, section parsing **System**: – Rules – 2 CRFs: one for numbers, one for names **Post-processing**: merging output, error correction [16]	0.9339	0.8823	0.90740
3	The University of Manchester	**System**: Combined outputs from two systems: – mDEID, an CRF-based system [5] – CliDEID, a 'data-driven' CRF system [8] **Post-processing**: combined outputs of two systems; kept longer of overlapping spans [6]	0.8888	0.8653	0.87690
4	University of Alabama at Birmingham	**Preprocessing**: tokenization, sentence and section detection **System**: multi-pass "sieve" system: – pattern matching – dictionary matching – Stanford CRF [10,3]	0.9162	0.8338	0.87310
5	Harbin Institute of Technology	**Pre-processing**: tokenization, sentence detection **System**: BI-LSTMs developed tags for each token **Post-processing**: a CRF layer identified most likely tag for each token [31]	0.8418	0.8728	0.85700
6	National Taitung University	**Pre-processing**: tokenization **System**: CRF [13]	0.7958	0.8501	0.82210
7	MITRE	System: MIST [4] with additional lexicons [2]	0.8552	0.762	0.80590
8	LIMSI-CNRS	**Pre-processing**: two text segmentations, token- and character-based **System**: 2 CRFs, one for each text segmentation **Post-processing**: merged outputs, prioritized character-based [12]	0.847	0.6963	0.76430
9	National Central University	**Pre-processing**: sentence detection, tokenization, POS tagging and chunking **System**: combination of rules and CRF **Post-processing**: dictionary matching, rules [25]	0.7892	0.779	0.75960
10	University of Utah #77	**Pre-processing**: sentence detection, tokenization, POS tagging, chunking **System**: Pipeline of dictionary matching, rules, and CRF [1]	0.8645	0.6132	0.71750

Stubbs et al. (2016). DEID Psych Notes: CEGS-NGRID. JBI.

biLSTM-CRF Architecture (Khin, 2018)

Figure 1: Deep Learning Architecture



Khin et al. (2018). Deep Learning DEID Patient Notes. Arxiv.

Systematic Review (Yogarajan; 2018, 2020): Open Research Questions

5.3. Research Questions

Although these de-identification competitions have provided a great platform to develop this area, some questions/areas are worth exploring.

(1) Considering the noticeable changes across the systems - is the hybrid system the best way to go? Could machine learning based system (for example, LSTM) outperform and be more adaptable to a new dataset?
(2) Is the use of rules and regular expressions crucial for rare, occurring PHIs? If so can we learn rules and regular expressions from data instead of handwritten rules (as seen in these competitions)?
(3) Apart from the noted fact that some PHIs, such as PROFESSION and LOCATION, perform poorly due to the low occurrence in training data, is there any other reason for this? Especially considering the use of handwritten rules and dictionary in the competitions, what are some other possible issues related to the poor F-measure? Any possible solutions to these problems?
(4) Adaptability to a new dataset is a noticeable issue with the de-identification systems. This is partly due to the lack of datasets. Could transfer learning be used to solve this problem? If so, is there any similar datasets available from domains such as law and crime science?
(5) Is there a way to obtain similar results to that of 2016 competitions with less complicated, reproducible systems?
(6) Considering the 95% on HIPAA regulations is still to be achieved across all PHIs - is there a need for stricter regulations, such as that introduced in these competitions? If so the re-identification implications will be an interesting avenue to explore.
(7) Systems developed using 2014 datasets after the competitions reported better performances for poorly performing PHIs. Is there a reason for this change? Would these systems do equally well if used for a different dataset? Can techniques from these systems be adopted to develop a more robust and better performing de-identification system?
(8) Would approaches such as word embeddings, which try and model the semantics of words, better be able to capture rarely occurring PHIs as opposed to syntactic approaches (given the lack of examples in the datasets to model the syntax of the rare PHIs)?

Yogarrajan et al. (2018). Survey of DEID of Longitudinal Clinical Narratives. Arxiv.

Selected Topic 4

EHRs for Randomized Controlled Trials

Why EHR Trials, Registry Trials, Trials with RCD, etc?

- Increasingly common design. Low cost. Large sample sizes. Pragmatic, generalizable.
- Methodological challenges: internally valid, selection effects, confounding? Ethics concerns?

EHR Trials (McCord et al)

KEY POINTS

- Randomized controlled trials (RCTs) using electronic health records (EHRs) can be directly integrated in routine care and allow large-scale and pragmatic trials with almost perfect generalizability.
- Currently, EHRs are mostly used in RCTs as part of the intervention, for example, when decision-support systems are evaluated as modifications or additions within the EHR infrastructure.
- Increasingly, EHR infrastructure is leveraged for patient recruitment or outcome assessment to support trials that assess conventional interventions.
- Randomized controlled trials within EHRs may address the limitations of traditional RCTs by increasing generalizability, reducing costs and time, expanding the research fields and allowing a democratization of research agendas.
- Important challenges include infrastructure costs, interoperability, standardization and quality of data, and ethical, privacy and data-security considerations.

Registry Trials (Thebane, et al)

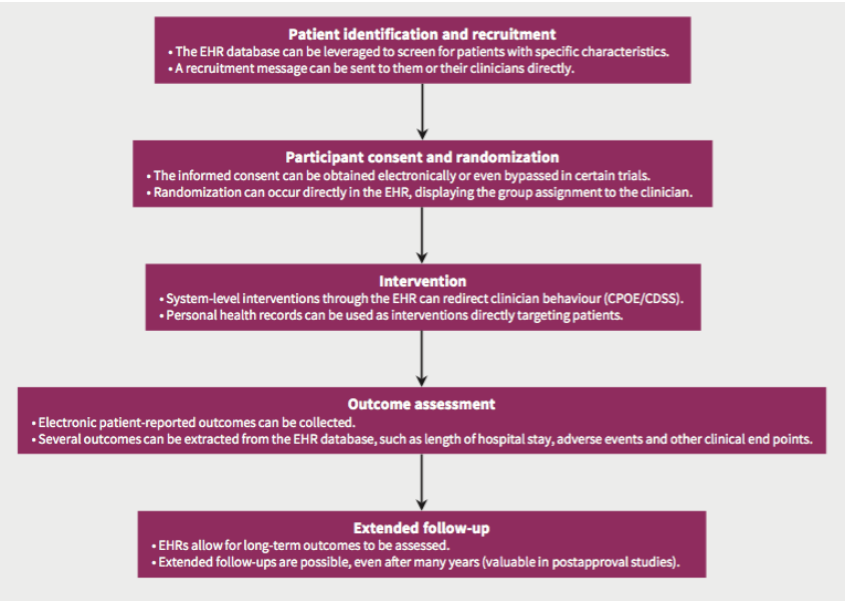
What is new?

- The application of registry-based randomized controlled trials has attracted increasing attention in health research to address comparative effectiveness research questions in real-world settings.
- When compared with conventional randomized effectiveness trials, the advantages of registry-based randomized controlled trials include low cost, enhanced generalizability of findings, rapid consecutive enrolment, and the potential completeness of follow-up for the reference population.
- Challenges of registry-based randomized controlled trials such as registry data quality, ethical issues and methodological challenges have to be taken into consideration before conducting such trials

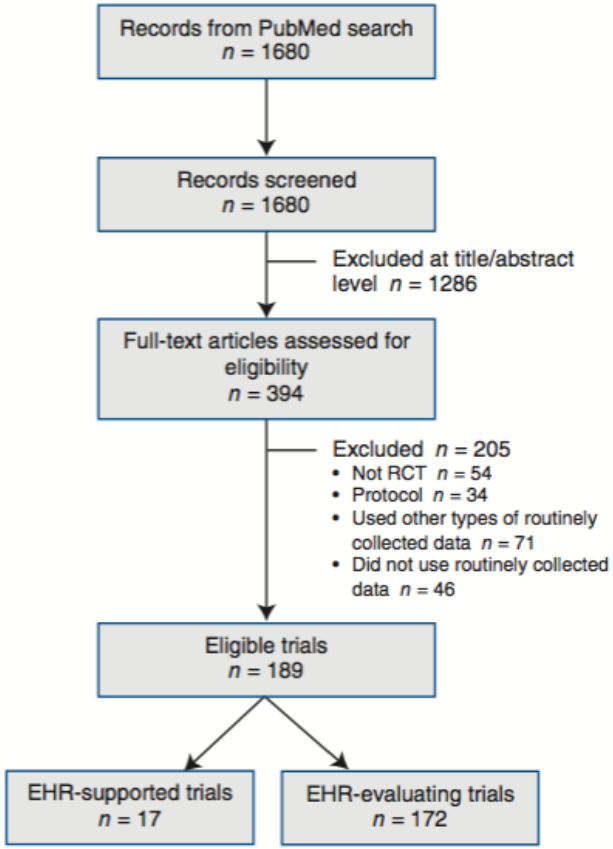
---

McCord et al. (2019). Using EHRs for Clinical Trials. CMAJ.
Thebane et al. (2016). Registry Based Randomized Controlled Trials. JCE.

Applications of EHR Trials



How EHRs Used in Trials



McCord et al. (2019). Current Use and Cost EHR Clinical Trials. CMAJ Open.
McCord et al. (2018). Using EHRs for Clinical Trials. CMAJ.

Characteristics of EHR Trials

Lensen et al (2020)

Table 1 Trial characteristics

Trial characteristic	Summary extraction: 2013–16, n (%) *n* = 69	Detailed extraction: 2017–18, n (%) *n* = 91	Total, n (%) *n* = 160
Purpose			
Screening	6 (9%)	10 (11%)	16 (10%)
Treatment	51 (74%)	65 (71%)	116 (73%)
Primary prevention	12 (17%)	16 (18%)	28 (18%)
Feasibility/Pilot trial			
Yes	6 (9%)	11 (12%)	17 (11%)
No	63 (91%)	80 (88%)	143 (89%)
Recruitment setting			
Primary care	23 (33%)	18 (20%)	41 (26%)
Secondary care	46 (67%)	73 (80%)	119 (74%)
Disease category			
Cancer	17 (25%)	30 (33%)	47 (29%)
Cardiovascular and stroke	21 (30%)	25 (27%)	46 (29%)
Pregnancy and childbirth	4 (6%)	5 (5%)	9 (6%)
Mental/neurological health	6 (9%)	6 (7%)	12 (8%)
Infection	5 (7%)	3 (3%)	8 (5%)
Endocrine and diabetes	1 (1%)	3 (3%)	4 (3%)
Inflammatory disorder	2 (3%)	3 (3%)	5 (3%)
Other	13 (19%)	16 (18%)	29 (18%)
Intervention			
Drug	38 (55%)	38 (42%)	76 (48%)
Surgical	5 (7%)	8 (9%)	13 (8%)
Other	26 (38%)	45 (49%)	71 (44%)
Primary outcome			
Survival related	33 (48%)	39 (43%)	72 (45%)
Other	36 (52%)	52 (57%)	88 (55%)
Randomisation			
Individual	61 (88%)	75 (82%)	136 (85%)
Cluster	8 (12%)	16 (18%)	24 (15%)
Trial size			
Median	1103	2611	1590
Range	41–170,432	53–6,000,000	41–6,000,000
1–500	21 (30%)	20 (22%)	41 (26%)
500–5000	31 (45%)	43 (47%)	74 (46%)
>5000	17 (25%)	25 (27%)	42 (26%)
Unclear	0 (0%)	3 (3%)	3 (2%)
International accrual			
Yes	13 (19%)	19 (21%)	32 (20%)
No (UK only)	56 (81%)	69 (76%)	125 (78%)
Unclear	0 (0%)	3 (3%)	3 (2%)
Coordinated by registered clinical trials unit			
Yes	40 (58%)	63 (69%)	103 (64%)
No	10 (14%)	12 (13%)	22 (14%)

McCord et al (2019)

Table 1: Characteristics of randomized controlled trials published in English between January 2000 and Sept. 13, 2017 that used electronic health records

Characteristic	No. (%) of trials* Overall *n* = 189	EHR-evaluating *n* = 172	EHR-supported *n* = 17
EHR for intervention	172 (91.0)	172 (100.0)	–
Computerized physician order entry system or clinical decision-support system	128 (67.7)	128 (74.4)	–
Telehealth	14 (7.4)	14 (8.1)	–
Personal health record	26 (13.8)	26 (15.1)	–
Electronic patient-reported outcomes	4 (2.1)	4 (2.3)	–
EHR for outcome measurement	158 (83.6)	143 (83.1)	15 (88)
EHR for patient recruitment	105 (55.6)	91 (52.9)	14 (82)
Country/region			
North America	153 (81.0)	140 (81.4)	13 (76)
United Kingdom	9 (4.8)	7 (4.1)	2 (12)
Continental Europe	15 (7.9)	14 (8.1)	1 (6)
Other†	12 (6.3)	11 (6.4)	1 (6)
Cluster RCT	64 (33.9)	61 (35.5)	3 (18)
Unit of randomization			
Clinician	49 (25.9)	46 (26.7)	3 (18)
Patient	76 (40.2)	65 (37.8)	11 (65)
Pharmacy	1 (0.5)	1 (0.6)	0 (0)
Practice/clinic	54 (28.6)	51 (29.6)	3 (18)
Unit/floor	9 (4.8)	9 (5.2)	0 (0)
Publication year, median (IQR)	2012 (2009–2014)	2012 (2009–2014)	2013 (2010–2013)
Sample size, median (IQR)			
Total	89 (24–732)	80 (22–513)	732 (73–2513)
Cluster RCTs excluded	239 (57–1187)	254 (60–1187)	900 (111–3075)
Cluster RCTs only	24 (12–47)	24 (12–52)	18 (12–24)
Blinding			
Open label	27 (14.3)	23 (13.4)	4 (24)
Single-blinded	19 (10.0)	18 (10.5)	1 (6)
Double-blinded	7 (3.7)	6 (3.5)	1 (6)
Outcome assessment blinding	35 (18.5)	30 (17.4)	5 (29)
Not reported	101 (53.4)	95 (55.2)	6 (35)

Note: EHR = electronic health record, IQR = interquartile range, RCT = randomized controlled trial.
*Except where noted otherwise.
†Includes China, Japan, Taiwan, Iran, India, Pakistan, Lebanon, Australia and Kenya.

McCord et al. (2019). Current Use and Cost EHR Clinical Trials. CMAJ Open.
Lensen et al. (2020). Access Routinely Collected Data for Trials. Trials.

Sampling from EHRs for RCTs

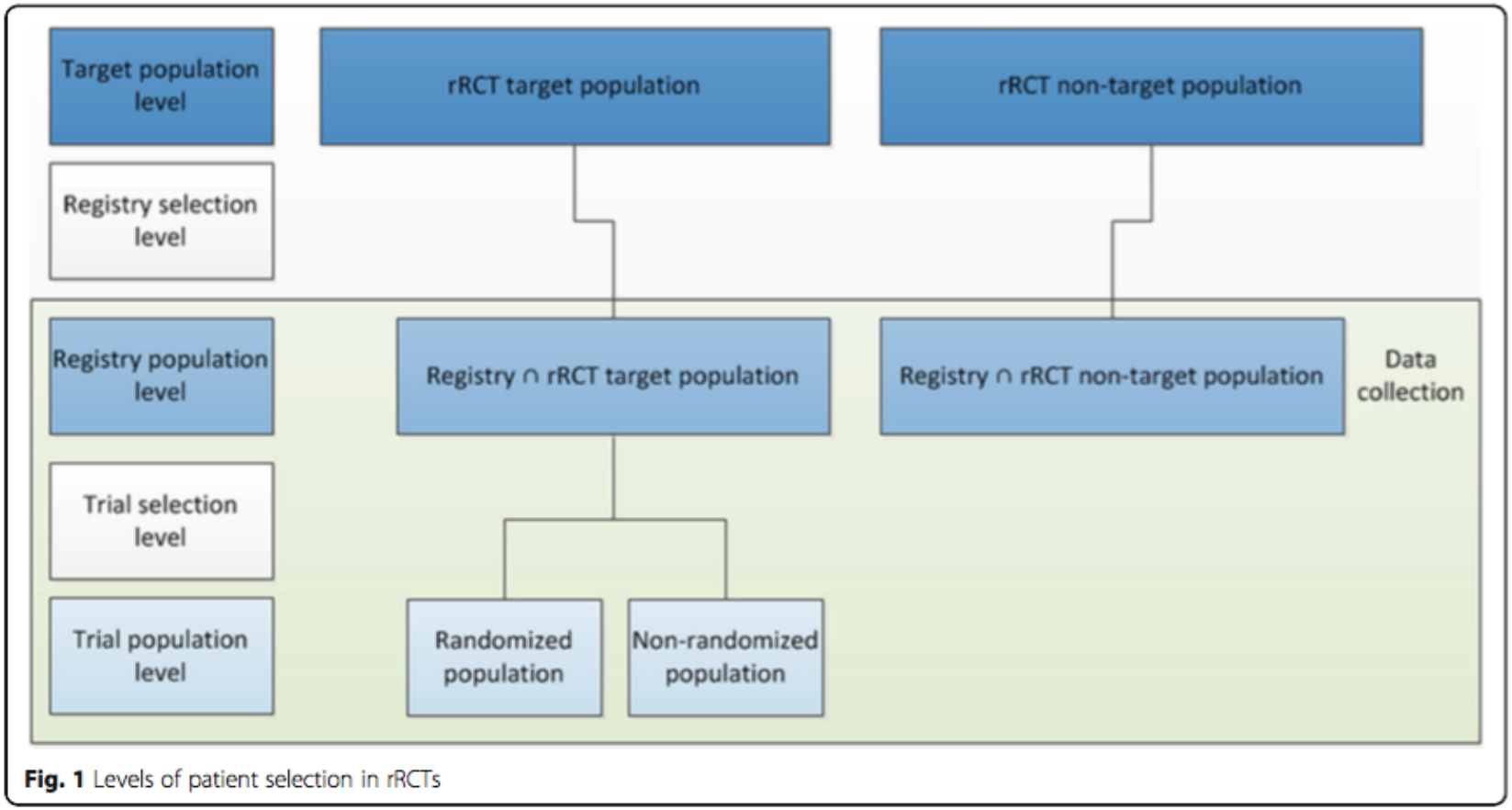


Fig. 1 Levels of patient selection in rRCTs

Lasch et al. (2020). Patient Selection in Registry Based Randomized Controlled Trials. Trials.

Strengths and Weaknesses of Registry Based RCTs

Table 2. Advantages and challenges of registry-based randomized controlled trials

Advantages	Challenges
Remarkably low cost	Registry data quality including the following:
Enhanced generalizability of findings	Definition, collection, and accuracy of baseline data gathered in registries may be various and questionable;
Rapid consecutive enrollment	Outcome data documented in registries may be subject to uncertainty;
Potential completeness of participant follow-up	Registries may have many missing data or fail to capture important prognostic factors.
	Ethical issues including the following:
	Screening registry participants for trial inclusion if they have not previously consented to records review;
	The potential need for formal informed consent for a treatment that is already being used in routine practice;
	Protecting the data and participant privacy;
	How to handle participant withdrawal from the trial or registry;
	How to coordinate the overlapping role of Data and Safety Monitoring Board in the trial with the role of registry executives.
	Methodological challenges including the following:
	Common confusion and controversies about the research question being addressed by the design;
	Ensuring the representativeness of study participants in recruitment;
	Research questions, study designs, and types of outcomes limited by quality and features of registry used.

Thebane et al. (2016). Registry Based Randomized Controlled Trials. JCE.

Advantages and Disadvantages of Registry Based RCTs

TABLE 1. Comparison of Study Designs

Study Design	Advantages	Disadvantages
Randomized controlled trial (RCT)	Minimize bias	Expensive
	Blinding possible	Requires time and scale
	Assess efficacy	Labor intensive
	Minimal to no missing data	Lack of generalizability
	Evidence "criterion standard"	Not feasible in all settings
		Frequently funded by Big Pharma
Observational studies using registry data	"Real world" representation	Data quality variable
	Inexpensive	Confounding and bias
	Time efficient	Hypothesis generating
	Large sample sizes (eg, detect rare events)	Lack of randomization
		Link to causality limited
Randomized registry trial (RRT)	Minimize bias	Data quality variable
	Blinding possible	Data collection limited to registry capabilities
	Inexpensive (eg, leverage established infrastructure)	Blinding challenging
	Relatively quick	Not feasible in all settings
	Large sample sizes	Pragmatic in nature
	Pragmatic in nature	May alter primary purpose of registry
	Greater generalizability	Relatively novel concept (ie, yet undefined standards, patient safety, adjudication, etc.)
	Flexible linkage to registry	Ethical and legal concerns
	Ability to retrospectively compare to nonenrolled patients	Data sharing issues

Liu et al. (2017). The Randomized Registry Trial. Annals Surgery.

Barriers, Pressing Questions and Possible Solutions in RCD Trials

Table 2 Barriers in the use of routinely collected health data for randomized controlled trials and options for improvement

General barriers or issues	Pressing questions	Possible solutions, actions and additional comments
Data		
• Availability • Management • Linkage • Accuracy • Validity	• Is the desired outcome variable or RCD source available? • Will it be possible to achieve the same data quality and accuracy with RCD as in traditional trials? • Is the data linkage and management feasible in institutions with limited IT infrastructure?	• A central register of databases available for clinical trial research would be helpful, ideally with details about data quality. • Establish core outcomes and structured outcome assessments in routine care • Create RCD trial guidelines and RCD source validation guidelines to help standardize their use and reduce sources of bias or uncertainty • Increase IT presence (particularly data analysts) to health research teams • The more RCD is sought out and used in research, the greater is its availability and differentiation.
Regulatory and ethics		
• Collecting and obtaining the data • Using and sharing the data	• What type of release must be given by the patients before their data can be collected or shared? • Is it ethical to use RCD without asking for their permission, even if their data are anonymized? • Can this data be considered of value and morally be sold? • How are concerns about privacy and informed consent approached (particularly in the context of population-wide trials or Zelen designs)? • Are data safety standards applied to RCD just as strictly as they are to traditional actively collected data? • Who is responsible for the safety of the data?	• Ethical guidelines specifically regarding the collection and dissemination of RCD should be developed. • Ethics and approval committees should deepen their knowledge of these novel ethical challenges. • Whereas personal data are collected daily from many sources (e.g., phone use), collection, storage, and dissemination of data related to health require more definite ethical oversight and greater transparency to the general public. • After safety issues are defined, researchers and stakeholders must ensure that data are safely handled, with full transparency of access.
Costs		
• Obtaining the data • Managing the data	• Will data collectors (e.g. health insurers) share their data? Freely or at a cost? • Is a constant increase in the generation of routine data really reducing the overall trial costs if the same institution collected the data in the first place? • When is the use of RCD cost-effective?	• The financial worth of health data is not defined or explored; empirical data are necessary to determine the cost of both producing and maintaining health data • Health data are already legally sold to many industries, and regulations/legislation must catch up with this aspect.
Novelty		
• Bureaucratic obstacles • Unawareness • Training to generate, collect, prepare, manage and analyze RCD for trials	• Will approval committees understand the implications of using RCD sources for clinical trials? • What are the challenges that can be expected bureaucratically because most submission templates do not assume the use of RCD and absence of patient contact? • Are data anonymization techniques clear? • What training is required to qualify individuals who generate, collect, prepare, and manage RCD for clinical trial research?	• Develop, in collaboration with approval committees, RCD-specific templates and submission forms, especially in such studies where no patient contact is foreseen and therefore speedy approval is desired. • Educate regarding data anonymization and confidentiality risks • Include the concept of using RCD for RCT in clinical research education and teaching • Create and use reporting guidance specifically for RCD-RCTs

McCord et al. (2018). Routinely Collected Data Trials. Trials.

Design and Methodological Considerations with EHR Trials

Experimental Designs in EHR Research

- Individual randomized RCTs (no so common).
- (Pragmatic) cluster RCTs.
- Stepped wedge designs (and other cross-over variants).
- Platform adaptive designs.
- Zelen designs and informed consent.

Causal Inference from Observational/Quasi-Experimental EHR Designs

- Propensity scores. Instrumental variable analysis. Mediation analysis.
- Causal Factor Analysis and SEM. Causal DAGs and NPSEM.

Hussey et al. (2007). Design Analysis Stepped Wedge Trials. Comtemp Clin Trials.
Zwarenstein et al. (2009). Making trials matter: pragmatic and explanatory trials. Trials.
Butler et al. (2018). Response-adaptive platform trials in primary care. BJGP.
Torgerson et al. (1998). What is Zelens Design. BMJ.
Campbell & Stanley. (1963). Experimental and Quasi-Experimental Designs for Research.
Rosenbaum. (2002). Observational Designs. Springer.
Hernan & Robins. (2019). Causal Inference: What If.
Bollen. (1989). Confirmatory Factor Analysis. Wiley.
Pearl. (2000). Causality.

Concluding Thoughts

Big Challenges with EHRs in Ontario (2020)

- Poor interoperability within silos of healthcare system?
- Non-existent interoperability across silos of healthcare system?
- Who owns the health data? Patients, providers, gov't, corp?

Big Open Questions...

- Why little/mixed evidence that EHRs improve patient health?
- Do EHRs negatively impact provider (mental) health?

Data as a Solution to Better Health?

- EHRs networks provide one data source for LHS.
- Need improved data partnerships. Clear data ownership.
- EHR data analytics involves unique inferential challenges.
- Constantly seek to improve data quality, design/methods.
- Integration research with practice to facilitate implementation.

Platt et al. (2020). The Learning Health System: A Review. JMIR.

Current Areas of Research Interest at the UTOPIAN EHR Network

- Design and evaluation of phenotyping studies.
- Deidentification of PHI from clinical text data.
- Non-negative matrix and tensor factorization for topic modelling.
- Primary care EHRs as platforms for controlled trials.
- Optical character recognition to extract data from scanned notes.
- DEID of OCR text data.
- Identifying, characterizing named entities in clinical text data.
- Integrating EHRs with BioBank data.
- PheWas studies.
- Mapping UTOPIAN EHR to OMOP CDM (OHDSI standard).

Thank You

christopher.meaney@utoronto.ca