A common pipeline for curating electronic health record data to enhance reproducibility of real-world evidence studies

Jue Hou
Jesse Gronsbell

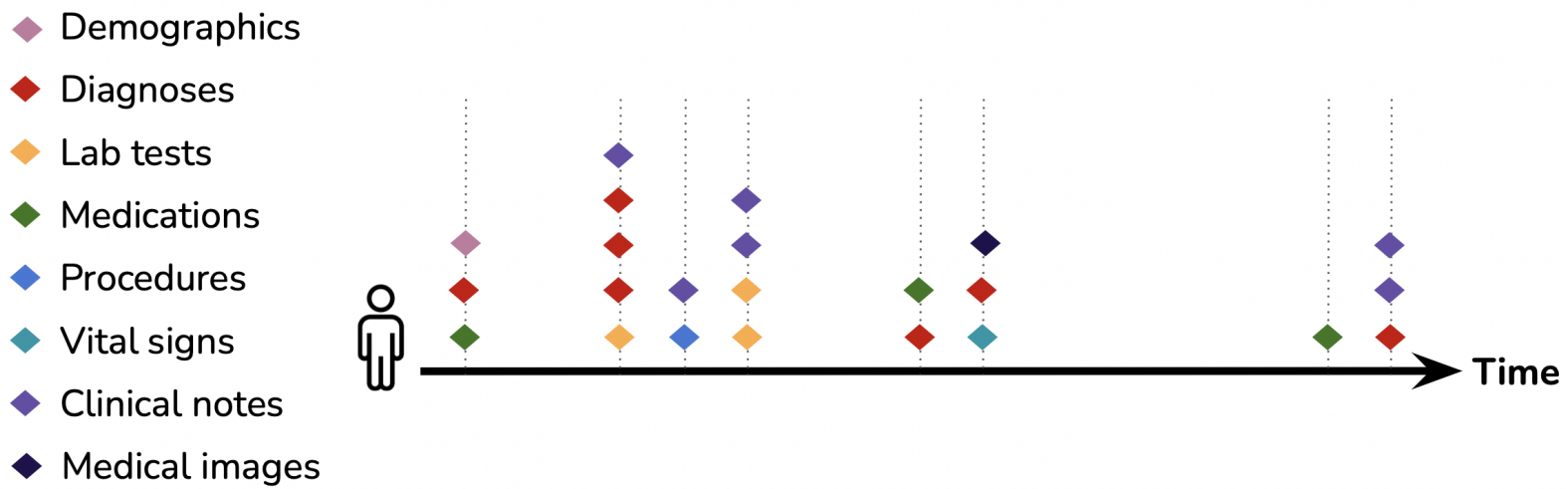
2023 Toronto Workshop on Reproducibility

Roadmap for today

- Background on electronic health records (EHRs)
- Reproducibility challenge
- Pipeline for curating EHRs for real-world evidence (RWE)

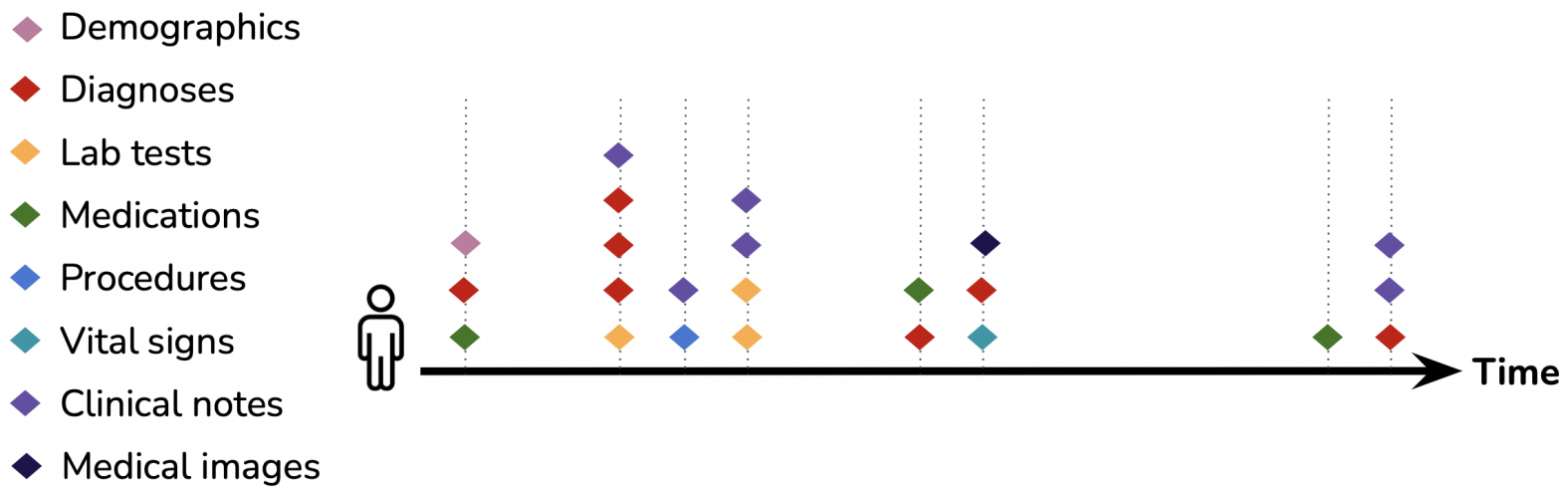
What is an Electronic Health Record (EHR)?

An electronic record of a patient's interactions with a healthcare system



What is an Electronic Health Record (EHR)?

An electronic record of a patient's interactions with a healthcare system



EHR data is a byproduct of clinical care

The blessing: EHR data is extensive

Big	Longitudinal records on large populations
Detailed	Information on numerous fields
Representative	Real-world patients
↑ **Available**	Increasing EHR adoption worldwide

The opportunity: Learn from EHR data

Healthcare professionals are beginning to tap the treasure trove of information locked in electronic health records to treat people in real time

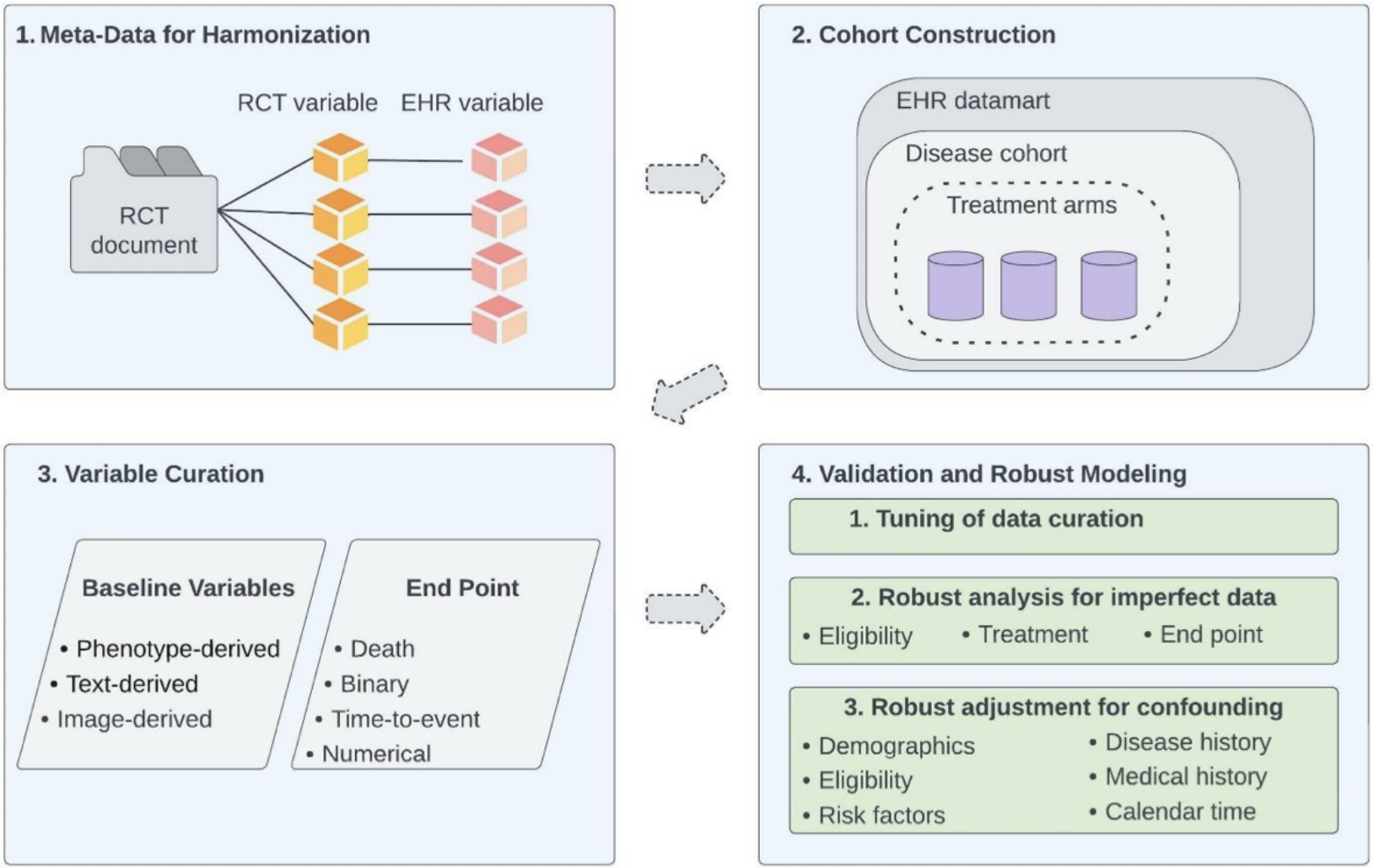
Data science at bedside: A "Green Button"

"a **green patients like mine button** as a tool in the EHR would both **support patient care decisions** in the absence of published evidence and, as a byproduct, quantify and **prioritize unanswered clinical questions** for EHR-enabled randomization at the point of care "

Longhurst et al 2014

Stanford Green Button Project

Data science at the bench: Real-world evidence



Hou J, Zhao R, Gronsbell J, et al. *Harnessing electronic health records for real-world evidence.* arXiv:2211.16609

The challenge: EHR data is not research ready

EHRs do not have readily available information on phenotypes

The challenge: EHR data is not research ready

EHRs do not have readily available information on phenotypes

Phenotypes: patient characteristics inferred from EHRs

- Presence of a disease
- Disease severity or subtype
- Time of disease onset
- Disease progression
- Treatment response
- ...

Phenotypes are the foundation of EHR research

- Presence of a disease
- Disease severity or subtype
- Time of disease onset
- Disease progression
- Treatment response
- ...

} **Identify and characterize the population of interest**

Example: Comparative effectiveness

Do patients with rheumatoid arthritis taking Infliximab or Adalimumab have a higher rate of symptomatic response?

Example: Comparative effectiveness

Do patients with rheumatoid arthritis taking Infliximab or Adalimumab have a higher rate of symptomatic response?

$\rightarrow$ **Cohort identification**

Which patients have rheumatoid arthritis?

Example: Comparative effectiveness

Do patients with rheumatoid arthritis taking Infliximab or Adalimumab have a higher rate of symptomatic response?

$\rightarrow$ **Cohort identification**

Which patients have rheumatoid arthritis?

$\rightarrow$ **Causal inference**

Which patients received treatment and who responded?

Example: Comparative effectiveness

Do patients with rheumatoid arthritis taking Infliximab or Adalimumab have a higher rate of symptomatic response?

→ **Cohort identification**

Which patients have rheumatoid arthritis?

→ **Causal inference**

Which patients received treatment and who responded?

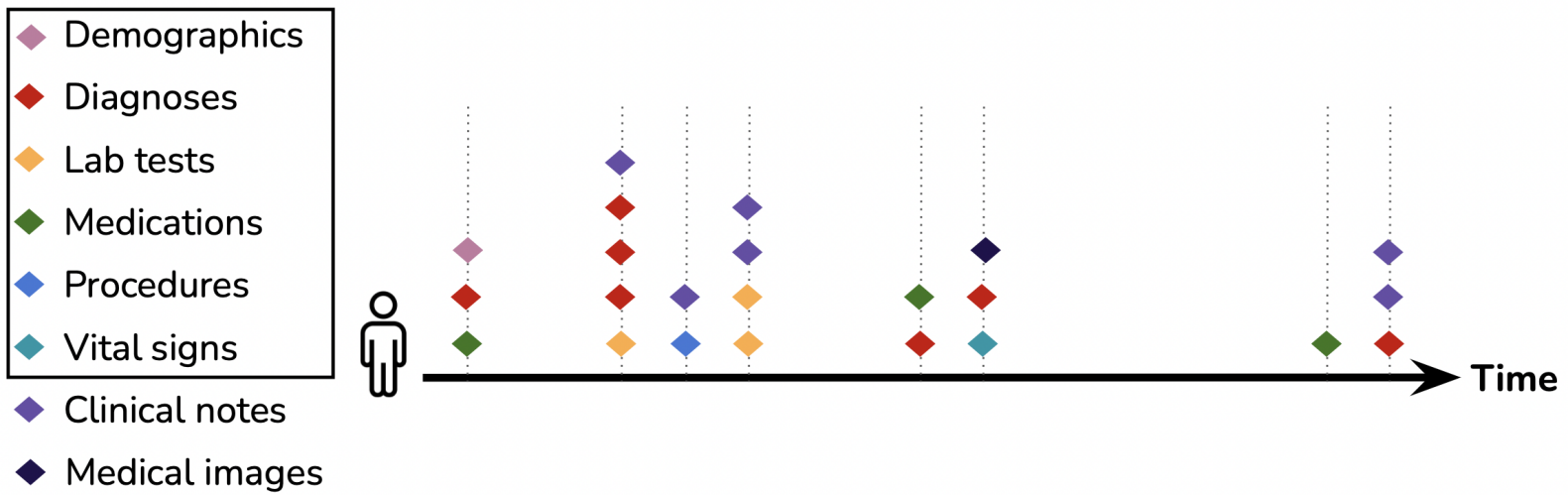
These are both phenotyping questions!

Why is phenotyping challenging?

"Health data is like crude oil. It is useless unless it is refined."

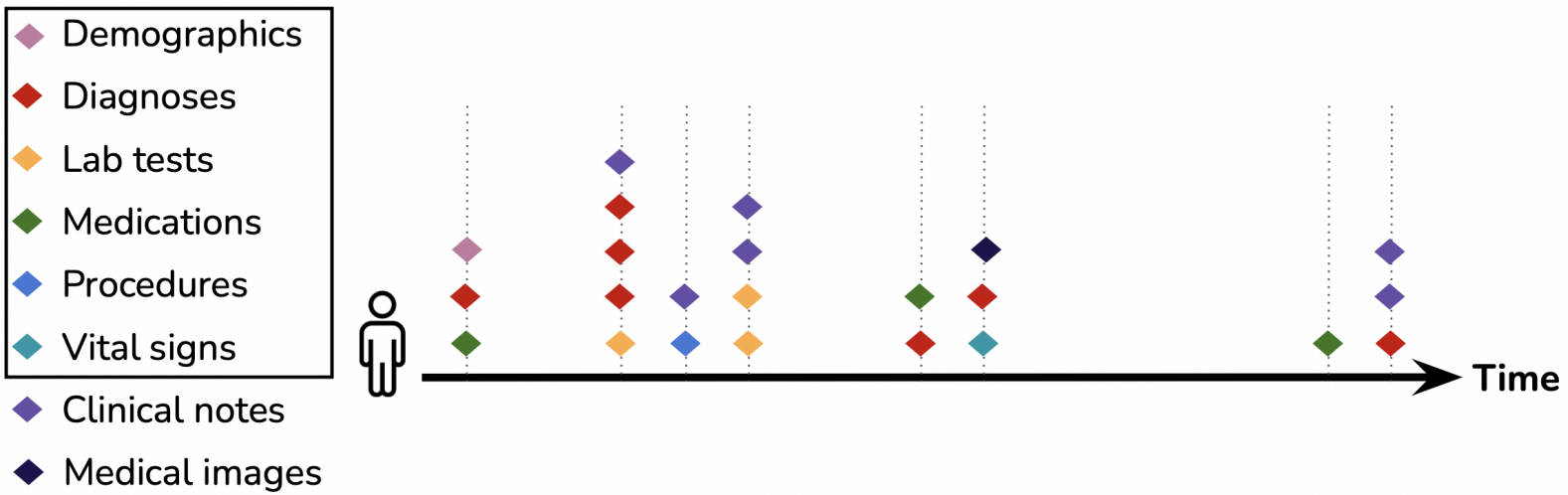
Leo Anthony Celi

The two flavors of EHR data



1. **Structured**: Easier to extract, but lacks context

The two flavors of EHR data

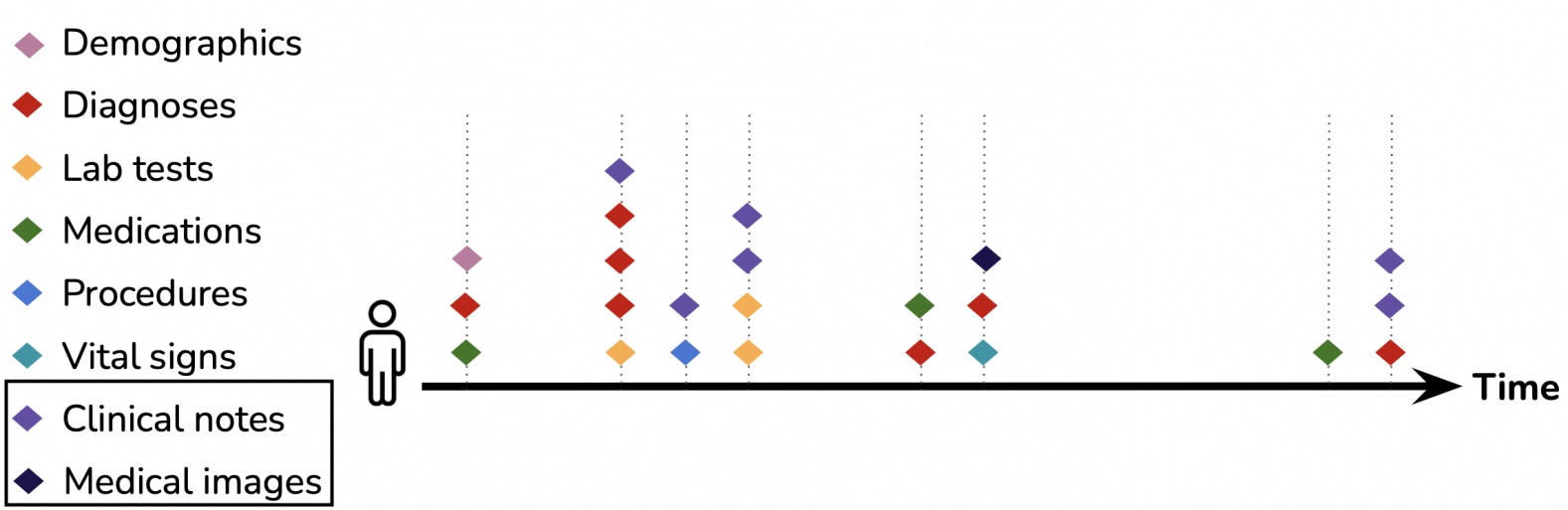


1. **Structured**: Easier to extract, but lacks context

e.g. diagnosis code $\not\to$ disease diagnosis

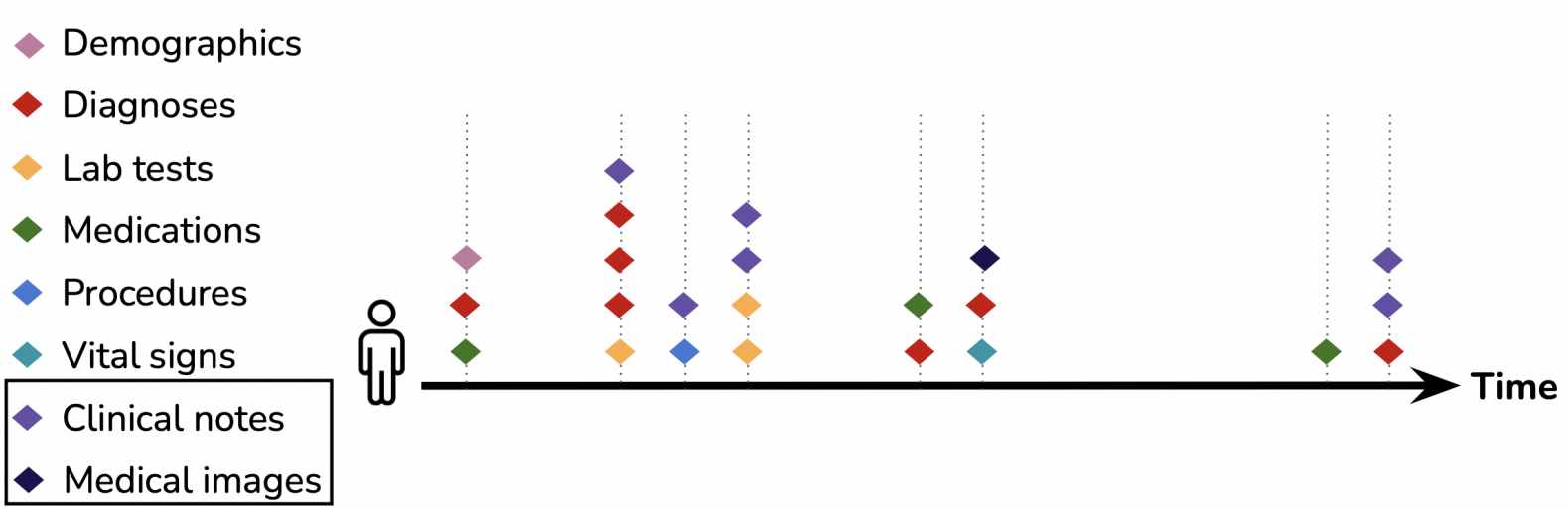
>3 diagnosis codes for rheumatoid arthritis
sensitivity $= 89\%$, positive predictive value $= 57\%$

The two flavors of EHR data



2. **Unstructured**: Rich information, but requires processing

The two flavors of EHR data

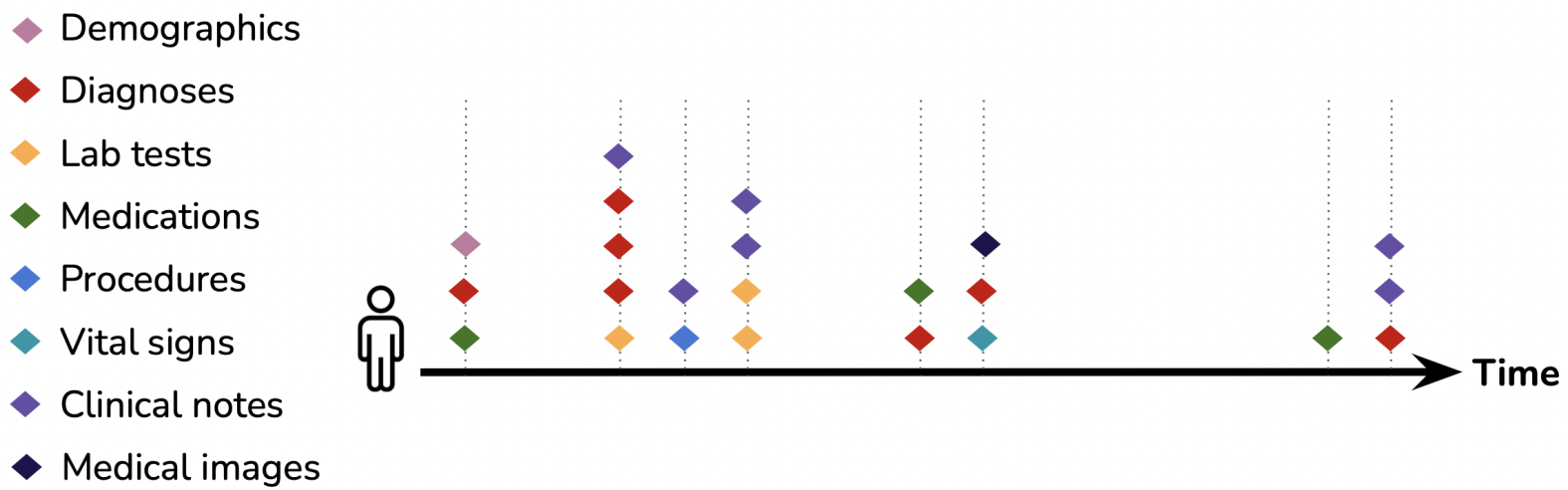


2. **Unstructured**: Rich information, but requires processing

e.g. clinical terms $\rightarrow$ medical concept

rheumatoid arthritis, ra, rha, r arthritis, ... = C000387

The two flavors of EHR data



$$\boxed{\text{phenotype} \approx \text{structured} + \text{unstructured data}}$$

The impact on reproducibility

Methods for enhancing the reproducibility of biomedical research findings using electronic health records

Spiros Denaxas[1,2*], Kenan Direk[1,2], Arturo Gonzalez-Izquierdo[1,2], Maria Pikoula[1,2], Aylin Cakiroglu[3], Jason Moore[4], Harry Hemingway[1,2] and Liam Smeeth[5]

"**only 5.1% of studies published** the entire set of controlled clinical terminology terms required to implement **the EHR-derived phenotypes used**"

The impact on reproducibility

Machine learning approaches for electronic health records phenotyping: a methodical review

Siyue Yang, Paul Varghese, Ellen Stephenson, Karen Tu, Jessica Gronsbell

Journal of the American Medical Informatics Association, Volume 30, Issue 2, February 2023, Pages 367–381, https://doi.org/10.1093/jamia/ocac216

"only 20% articles released their analytic code"

Example: Real-world evidence (RWE)

Reproducibility of real-world evidence studies using clinical practice data to inform regulatory and coverage decisions

Shirley V. Wang ✉, Sushama Kattinakere Sreedhara, Sebastian Schneeweiss & REPEAT Initiative

Nature Communications **13**, Article number: 5126 (2022) | Cite this article

"unambiguous communication about **the complex data processing**, design and analytic choices involved in RWE studies improves understanding of **the reproducibility of evidence**"

Why Real-world Evidence (RWE)?

- Initiation from the legislation: accelerate the use of RWE in the discovery, development and delivery of medical treatments.

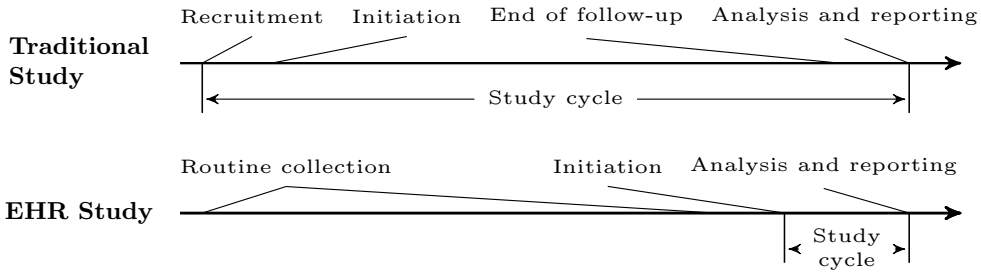
  114th Congress (1996). 21st century cures act

- Framework specified by regulatory agency (FDA): including approvals for new indications and post-approval requirements.

  US Food and Drug Administration (2018). Framework for FDA's real-world evidence program

  US Food and Drug Administration (2020). Real world evidence - from safety to a potential tool for advancing innovative ways to develop new medical therapies

Why Real-world Evidence (RWE)?



- Trials not always possible: generalisability, adaptibility, feasibility, ethics, and power.
- RWE had a record in regulatory decision making.

Franklin, J. M., Liaw, K.-L., Iyasu, S., Critchlow, C. W., and Dreyer, N. A. (2021a). Real-world evidence to support regulatory decision making: New or expanded medical product indications.

Pharmacoepidemiology and Drug Safety, 30(6):685–693

- The trending: use RWD (Data) to supplement trials.

Rogers, J. R., Lee, J., Zhou, Z., Cheung, Y. K., Hripcsak, G., and Weng, C. (2020). Contemporary use of real-world data for clinical trial conduct in the United States: a scoping review.

Journal of the American Medical Informatics Association, 28(1):144–154

RWE: the Path toward Reproducibility

- FDA guideline calls for transparency and reproducibility

  US Food and Drug Administration (2018). Framework for FDA's real-world evidence program

- A first step: Pre-registration with pre-specified analysis

  Franklin, J. M., Pawar, A., Martin, D., Glynn, R. J., Levenson, M., Temple, R., and Schneeweiss, S. (2020). Nonrandomized real-world evidence to support regulatory decision making: Process for a randomized trial replication project.

  Clinical Pharmacology & Therapeutics, 107(4):817–826

  Franklin, J. M., Patorno, E., Desai, R. J., Glynn, R. J., Martin, D., Quinto, K., Pawar, A., Bessette, L. G., Lee, H., Garry, E. M., Gautam, N., and Schneeweiss, S. (2021b). Emulating randomized clinical trials with nonrandomized real-world evidence studies.

  Circulation, 143(10):1002–1013

- Missing piece: the data!

A tutorial for the community

- Creation of analysis data lacks
  - Prespecified data creation plan;
  - Validation of data quality.

  Hard to find a single paper detailing the process!
- A tutorial

  Hou, J., Zhao, R., Gronsbell, J., Beaulieu-Jones, B. K., Webber, G., Jemielita, T., Wan, S., Hong, C., Lin, Y., Cai, T., Wen, J., Panickan, V. A., Bonzel, C.-L., Liaw, K.-L., Liao, K. P., and Cai, T. (2022).

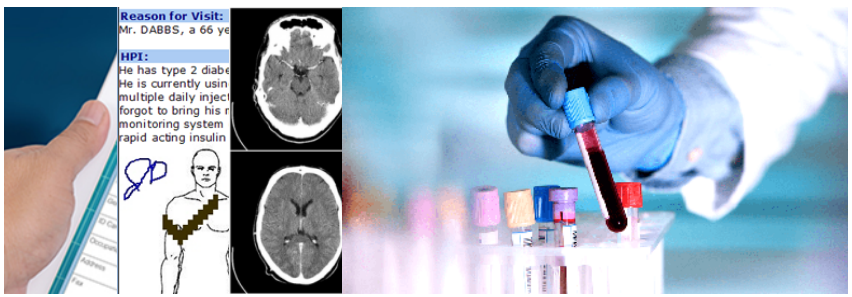
  Harnessing electronic health records for real-world evidence

  - Standard data creation protocol for future design;
  - State-of-art computational tools for best quality;
  - Validation strategy for quality assurance;
  - Robust causal modeling for faithful RWE.

Data harmonization

- Scattered information without precise (1-1) indicator.
  - diagnosis, combination therapies and disease progression.

Longitudinal EHR data in all formats → **Analysis data (traditional reporting from)**



Patient ID: Center ID: Date:
Name of the Patient: Age (years): __ Sex: M F
Contact Number: Address:

History of	Yes	Duration
Smoking		
Alcohol		
Burning in feet		
Amputation		
Heart attack/ stroke		
Diabetes		
Hypertension		
Dyslipidemia		
On ACE-I		
On ARBs		
On statins		

Family history of	Yes	No
Diabetes mellitus		
Kidney disease		
High blood pressure		
Heart disease		

Personal Details	Details
Height (cm)	
Weight (kg)	
Waist circumference (cm)	
BMI	
Education	

- Organization of EHR data
  1. Grouping of structured codes;
  2. Sequencing of free text data;
  3. Variable specific tools for language or image data.
- Mapping design to EHR
  1. Curate medical concepts from protocol/article;
  2. Network between medical concepts and EHR features;
  3. Expanded mapping for poorly structured variables.

Feature curation

- Tools classified by types
  - Binary status: phenotyping.
  - Longitudinal/time-to-event: incidence phenotyping.
  - Text derived: natural language processing.
  - Image derived: AI.
- Roles in data creation
  1. Construction of disease cohort;
  2. Identification of intervention arms;
  3. Extraction of key eligibility/characteristics;
  4. Curation of endpoints.

Validation and causal modeling

- Validation by manual abstraction.
  - Assessment data quality;
  - Calibrate initial extraction;
  - Robust statistical analysis.
- Causal modeling.
  - Robust to imperfect data;
  - Data driven confounding capture through high-dimensional summary of EHR features and semantic embeddings;
  - Account for temporal trends and population shift.

  Data must be prepared to facilitate the downstream modeling!

Future directions

- More tools for currently unavailable features
  1. Disease activity, progression and severity;
  2. Other mobility, cognitive measures;
  3. Life styles;
  4. Social determinant of health.
- Harmonization across healthcare systems.
- Stable personalized treatment guideline.