Weill Cornell Medicine

Weill Cornell Medicine
Institute of Artificial Intelligence for Digital Health

KDD2023 AUGUST 6 - 10 LONG BEACH, CA

KDD 2023 Tutorial – LS09
Mining Electronic Health Records for Real-World Evidence

Tuesday, August 8th, 10:00 am-13:00 pm PDT, Room 202A

Chengxi Zang, PhD, Weishen Pan, PhD, & Fei Wang, PhD
Department of Population Health Sciences
Institute of Artificial Intelligence for Digital Health (AIDH)
Weill Cornell Medicine, Cornell University

www.calvinzang.com/ehr4rwe_kdd2023.html

Outline

- Generating Real-World Evidence for Understanding Long COVID
- Advancements in Risk Prediction using EHRs
- Discussion & QA

Weill Cornell Medicine

Weill Cornell Medicine
Institute of Artificial Intelligence for Digital Health

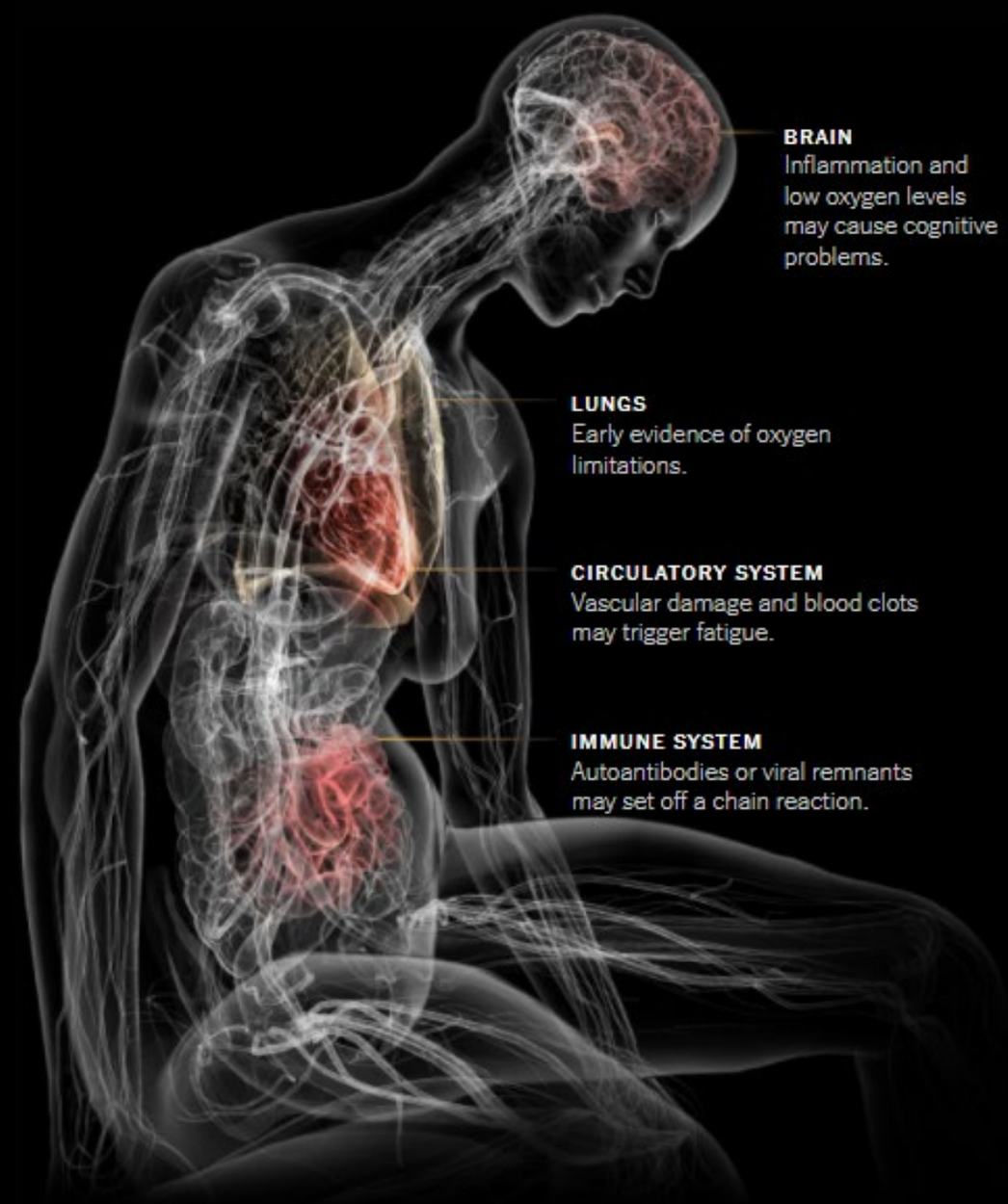
KDD2023 AUGUST 6 - 10 LONG BEACH, CA

Part-1: Generating Real-World Evidence for Understanding Long COVID

Tuesday, August 8th, 10:00 am-13:00 pm PDT, Room 202A

Chengxi Zang, PhD, Weishen Pan, PhD, & Fei Wang, PhD
Instructor @ Department of Population Health Sciences
Institute of Artificial Intelligence for Digital Health (AIDH)
Weill Cornell Medicine, Cornell University
www.calvinzang.com/ehr4rwe_kdd2023.html
chz4001@med.cornell.edu
www.calvinzang.com
@calvin_zcx

Long COVID



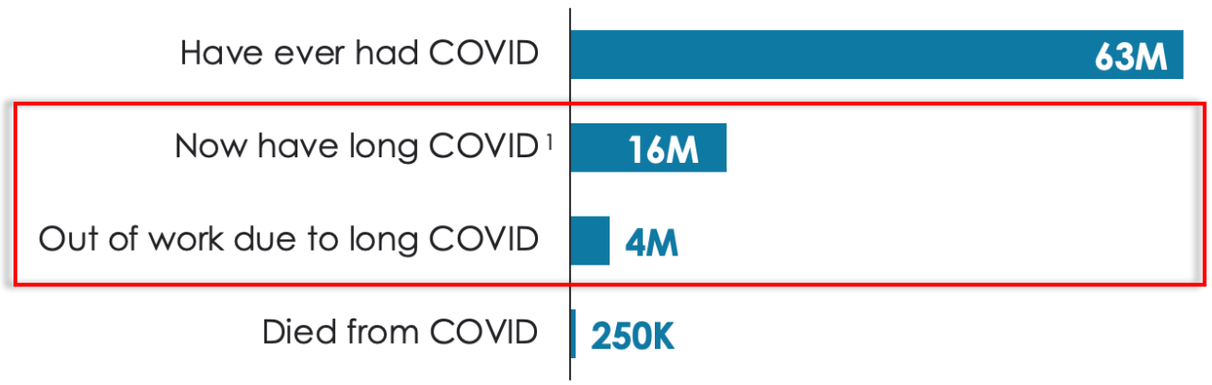
Millions of people continue to suffer from exhaustion, cognitive problems and other long-lasting symptoms after a coronavirus infection. The exact causes of the illness, known as long Covid, are not known. But new research offers clues, describing the toll the illness takes on the body and why it can be so debilitating.

nytimes

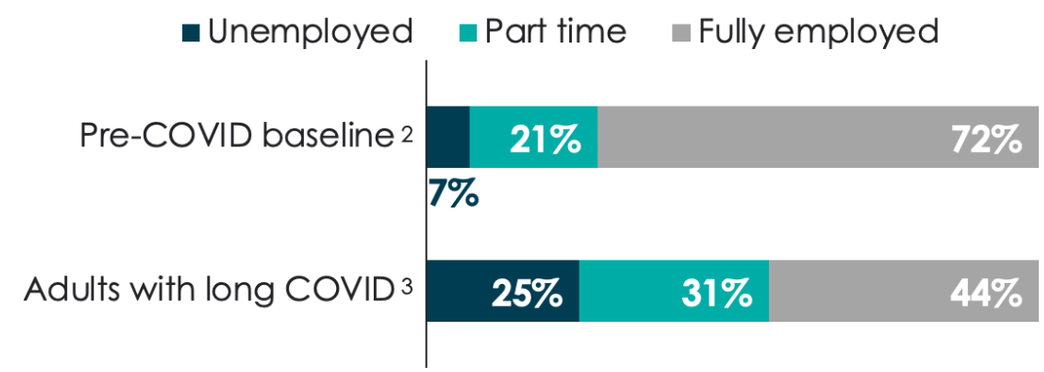
Long COVID Keeping Millions Out of Workforce

Up to 40 Percent of Job Openings Unfilled Because of Long COVID

Estimated Prevalence of US COVID Infections, Adults Ages 18-64, Oct. 2022

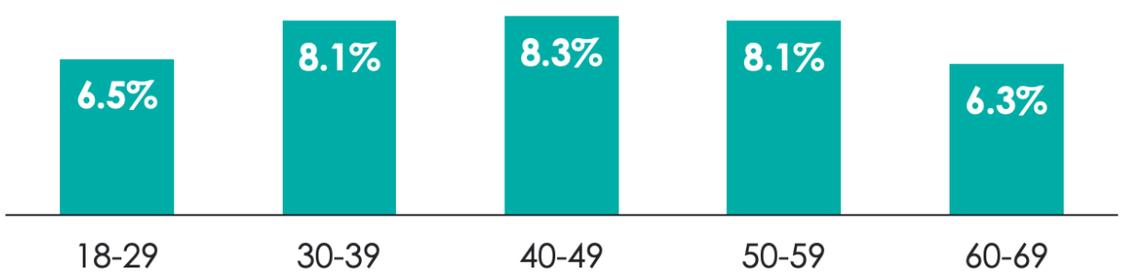


Comparing Employment Profiles of All Adults Pre-COVID to Adults with Long COVID



Percentage of Working Age Adults Who Currently Have Long COVID, by Age Group

n = 41,415; Oct. 2022



Estimated Impact of Long COVID on the Economy

40%

Percentage of the **10M current job openings potentially unfilled** because of long COVID

$235B

Total annual wages **lost due to long COVID**

1. Long COVID diagnosis determined by self-reported symptoms to US Census Bureau's Household Pulse Survey.
2. Based on 2019 Current Population Survey profile of US workforce.
3. Based on two studies (n=3,762, May 2020; n=3,296, Nov. 2021) of adults in the workforce diagnosed with long COVID.

Source: Centers for Disease Control and Prevention. "COVID Data Tracker." 27 Oct. 2022; Bach, Katie. "New data shows long Covid is keeping as many as 4 million people out of work." *Brookings.* 24 Aug. 2022; Burns, Alice. "What are the Implications of Long COVID for Employment and Health Coverage?." *Kaiser Family Foundation.* 1 Aug. 2022; Bureau of Labor Statistics. "Job Openings and Labor Turnover Survey." 4 Oct. 2022; Gist Healthcare analysis.

THE NIH DIRECTOR

February 23, 2021

NIH launches new initiative to study "Long COVID"

I write to announce a major new NIH initiative to identify the causes and ultimately the means of prevention and treatment of individuals who have been sickened by COVID-19, but don't recover fully over a period of a few weeks. Large numbers of patients who have been infected with SARS-CoV-2 continue to experience a constellation of symptoms long past the time that they've recovered from the initial stages of COVID-19 illness. Often referred to as "Long COVID", these symptoms, which can include fatigue, shortness of breath, "brain fog", sleep disorders, fevers, gastrointestinal symptoms, anxiety, and depression, can persist for months and can range from mild to incapacitating. In some cases, new symptoms arise well after the time of infection or evolve over time. In December, NIH held a workshop to summarize what is known about these patients who do not fully recover and identify key gaps in our knowledge about the effects of COVID-19 after the initial stages of infection. In January, I shared the results from the largest global study of these emerging symptoms. While still being defined, these effects can be collectively referred to as Post-Acute Sequelae of SARS-CoV-2 infection (PASC). We do not know yet the magnitude of the problem, but given the number of individuals of all ages who have been or will be infected with SARS-CoV-2, the coronavirus that causes COVID-19, the public health impact could be profound.

In December, Congress provided $1.15 billion in funding over four years for NIH to support research into the prolonged health consequences of SARS-CoV-2 infection. A diverse team of experts from across the

RECOVER: Researching COVID to Enhance Recovery

The National Institutes of Health (NIH) created the RECOVER Initiative to learn about the long-term effects of COVID.

The goal of RECOVER is to rapidly improve our understanding of and ability to predict, treat, and prevent PASC (post-acute sequelae of SARS-CoV-2), including Long COVID.

LEARN MORE ABOUT LONG COVID

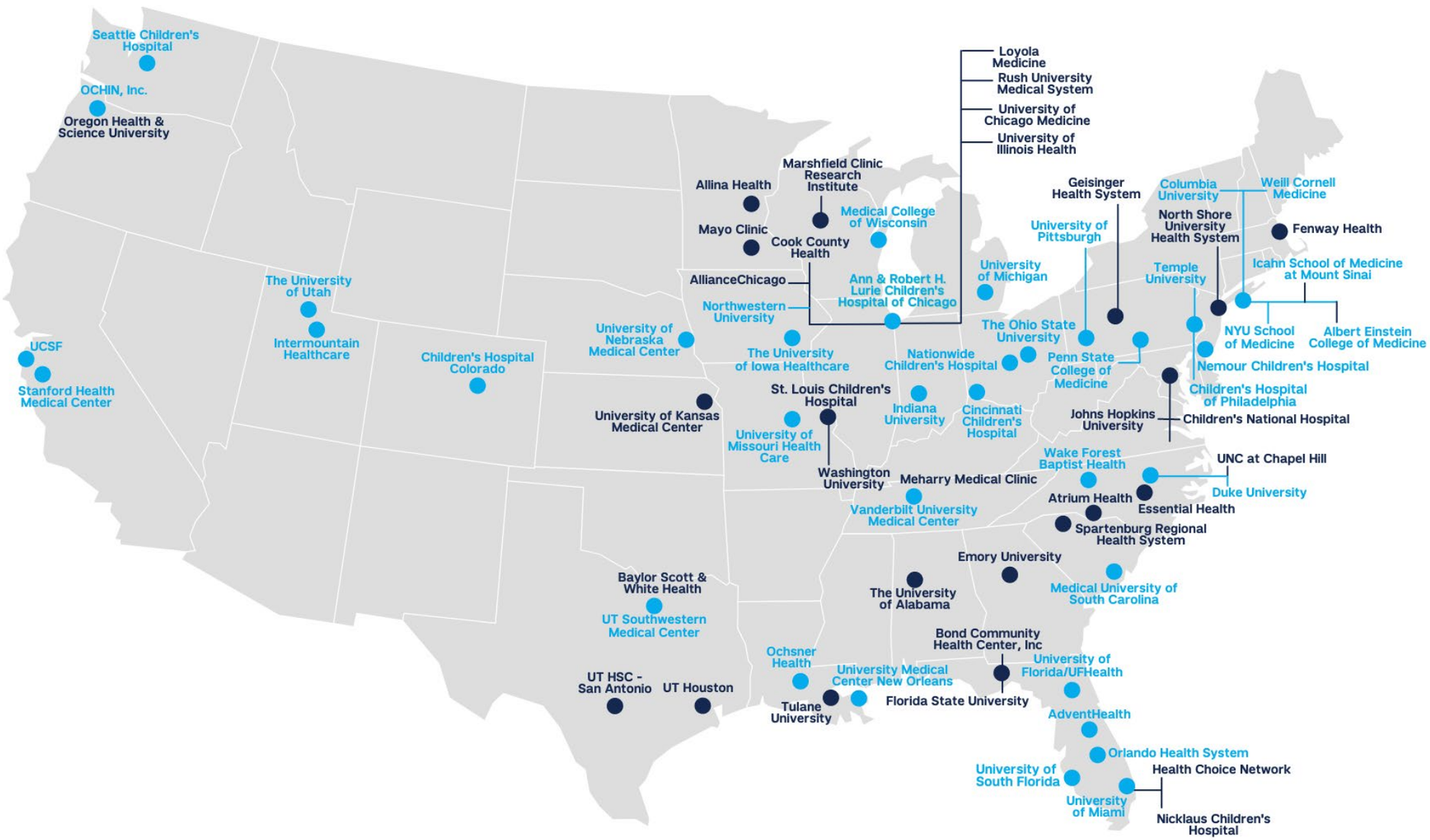
People are joining the search for answers.

RECOVER is a first of its kind research initiative created specifically to address the widespread and diverse impacts of Long COVID. Thousands of children and adults - including pregnant people - have joined RECOVER studies.

Leveraging EHR/RWD to Understand Long COVID

pcornet® The National Patient-Centered Clinical Research Network

RECOVER Researching COVID to Enhance Recovery
An Initiative Funded by the National Institutes of Health



15B+ rows of data

- Electronic health records
  - structured
  - unstructured
- Public payor data
- Exposome data
  - race/ethnicity
  - socio-economic
  - environmental
- Vaccine data

- PCORnet Sites (n=65)
- RECOVER Data Enriched Sites (n=41 sites)

PCORnet Adult Research

Leveraging EHR/RWD to Understand Long COVID

Predict

Geographic, demographic, socioeconomic disparities, Examine risk factors

Treat

Characterize treatments and patterns of therapeutic use, Therapeutic effectiveness

Prevent

Vaccination linkage and quality improvement
Vaccine effectiveness

Define & Detect

Phenotype development, refinement, validation, Characterize PASC

- Epidemiology & Health Services Research
- Machine Learning & Artificial Intelligence
- Queries

Recent Sample of Papers

pcornet® The National Patient-Centered Clinical Research Network

RECOVER Researching COVID to Enhance Recovery
An Initiative Funded by the National Institutes of Health

nature communications

Article

https://doi.org/10.1038/s41467-023-37653-z

Data-driven analysis to understand long COVID using electronic health records from the RECOVER initiative

Received: 8 June 2022

Accepted: 24 March 2023

Published online: 07 April 2023

Check for updates

Chengxi Zang[1], Yongkang Zhang[1], Jie Xu[2], Jiang Bian[2], Dmitry Morozyuk[1], Edward J. Schenck[3], Dhruv Khullar[1], Anna S. Nordvig[4], Elizabeth A. Shenkman[2], Russell L. Rothman[5], Jason P. Block[6], Kristin Lyman[7], Mark G. Weiner[1], Thomas W. Carton[7], Fei Wang[1] & Rainu Kaushal[1]

Recent studies have investigated post-acute sequelae of SARS-CoV-2 infection (PASC, or long COVID) using real-world patient data such as electronic health records (EHR). Prior studies have typically been conducted on patient cohorts with specific patient populations which makes their generalizability unclear. This study aims to characterize PASC using the EHR data warehouses from two large Patient-Centered Clinical Research Networks (PCORnet), INSIGHT and OneFlorida+, which include 11 million patients in New York City (NYC) area and 16.8 million patients in Florida respectively. With a high-throughput screening pipeline based on propensity score and inverse probability of treatment weighting, we identified a broad list of diagnoses and medications which exhibited significantly higher incidence risk for patients 30–180 days after the laboratory-confirmed SARS-CoV-2 infection compared to non-infected patients. We identified more PASC diagnoses in NYC than in Florida regarding our screening criteria, and conditions including dementia, hair loss, pressure ulcers, pulmonary fibrosis, dyspnea, pulmonary embolism, chest pain, abnormal heartbeat, malaise, and fatigue, were replicated across both cohorts. Our analyses highlight potentially heterogeneous risks of PASC in different populations.

The global COVID-19 pandemic from late 2019 has led to more than 620 million infections and 6.5 million deaths as of Oct 17, 2022[1]. Growing scientific and clinical evidence has demonstrated potential post-acute and long-term effects of SARS-CoV-2 infection in multiple organ systems[2], including cardiovascular[3], mental health[4], neurological[5], and metabolic[6] among other systems. Recently, several retrospective observational cohort analyses have described post-acute sequelae of SARS-CoV-2 infection (PASC) using real-world patient data[7–9]. These studies typically start with a predefined list of PASC symptoms and signs and then contrast their incidence risk or burden in SARS-CoV-2 infected patients versus non-infected controls. Different analytical pipelines have been utilized, such as causal inference[7], regression analysis[10], and network analysis[11]. There are two major challenges to these existing studies. First, the disease etiology and

[1]Department of Population Health Sciences, Weill Cornell Medicine, New York, NY, USA. [2]Department of Health Outcomes Biomedical Informatics, University of Florida, Gainesville, FL, USA. [3]Department of Medicine, Division of Pulmonary and Critical Care Medicine, Weill Cornell Medicine, New York, NY, USA. [4]Department of Neurology, Weill Cornell Medicine, New York, NY, USA. [5]Center for Health Services Research, Vanderbilt University Medical Center, Nashville, TN, USA. [6]Department of Population Medicine, Harvard Pilgrim Health Care Institute, Harvard Medical School, Boston, MA, USA. [7]Louisiana Public Health Institute, New Orleans, LA, USA. e-mail: few2001@med.cornell.edu

nature medicine

Article

https://doi.org/10.1038/s41591-022-02116-3

Data-driven identification of post-acute SARS-CoV-2 infection subphenotypes

Received: 8 June 2022

Accepted: 2 November 2022

Published online: 1 December 2022

Check for updates

Hao Zhang[1], Chengxi Zang[1], Zhenxing Xu[1], Yongkang Zhang[1], Jie Xu[2], Jiang Bian[2], Dmitry Morozyuk[1], Dhruv Khullar[1], Yiye Zhang[1], Anna S. Nordvig[3], Edward J. Schenck[4], Elizabeth A. Shenkman[2], Russell L. Rothman[5], Jason P. Block[6], Kristin Lyman[7], Mark G. Weiner[1], Thomas W. Carton[7], Fei Wang[1] & Rainu Kaushal[1]

The post-acute sequelae of SARS-CoV-2 infection (PASC) refers to a broad spectrum of symptoms and signs that are persistent, exacerbated or newly incident in the period after acute SARS-CoV-2 infection. Most studies have examined these conditions individually without providing evidence on co-occurring conditions. In this study, we leveraged the electronic health record data of two large cohorts, INSIGHT and OneFlorida+, from the national Patient-Centered Clinical Research Network. We created a development cohort from INSIGHT and a validation cohort from OneFlorida+ including 20,881 and 13,724 patients, respectively, who were SARS-CoV-2 infected, and we investigated their newly incident diagnoses 30–180 days after a documented SARS-CoV-2 infection. Through machine learning analysis of over 137 symptoms and conditions, we identified four reproducible PASC subphenotypes, dominated by cardiac and renal (including 33.75% and 25.43% of the patients in the development and validation cohorts); respiratory, sleep and anxiety (32.75% and 38.48%); musculoskeletal and nervous system (23.37% and 23.35%); and digestive and respiratory system (10.14% and 12.74%) sequelae. These subphenotypes were associated with distinct patient demographics, underlying conditions before SARS-CoV-2 infection and acute infection phase severity. Our study provides insights into the heterogeneity of PASC and may inform stratified decision-making in the management of PASC conditions.

The ongoing global pandemic of Coronavirus Disease 2019 (COVID-19) caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) infection has impacted hundreds of millions of people's lives. Existing studies have provided evidence that many symptoms and signs could be persistent, exacerbated or newly present after the acute phase of SARS-CoV-2 infection, referred to as post-acute sequelae of SARS-CoV-2 infection (PASC)[1,2], which involve multiple organ systems, including cardiovascular[3], mental[4], metabolic[5], renal[6] and others. There have been various ongoing efforts into investigating the underlying biological mechanisms of PASC[7–9], which have typically been conducted in small patient cohorts. Large-scale clinical observational cohort studies can provide useful insights into PASC that may help develop

[1]Department of Population Health Sciences, Weill Cornell Medicine, New York, NY, USA. [2]Department of Health Outcomes Biomedical Informatics, University of Florida, Gainesville, FL, USA. [3]Department of Neurology, Weill Cornell Medicine, New York, NY, USA. [4]Department of Medicine, Division of Pulmonary and Critical Care Medicine, Weill Cornell Medicine, New York, NY, USA. [5]Center for Health Services Research, Vanderbilt University Medical Center, Nashville, TN, USA. [6]Department of Population Medicine, Harvard Pilgrim Health Care Institute, Harvard Medical School, Boston, MA, USA. [7]Louisiana Public Health Institute, New Orleans, LA, USA. e-mail: few2001@med.cornell.edu

Outline

1. Backgrounds

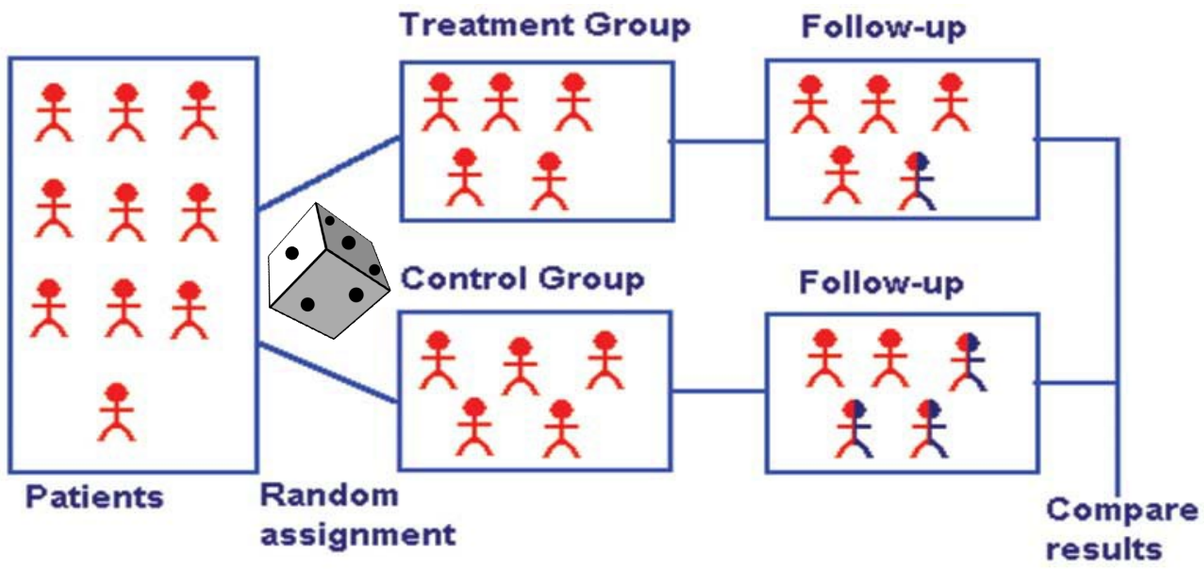
2. Key Concepts & Causal Basics

   - RCT, RWD/E, Trial Emulation, Causal Inference, etc.

3. Applications & Beyond

4. Conclusions

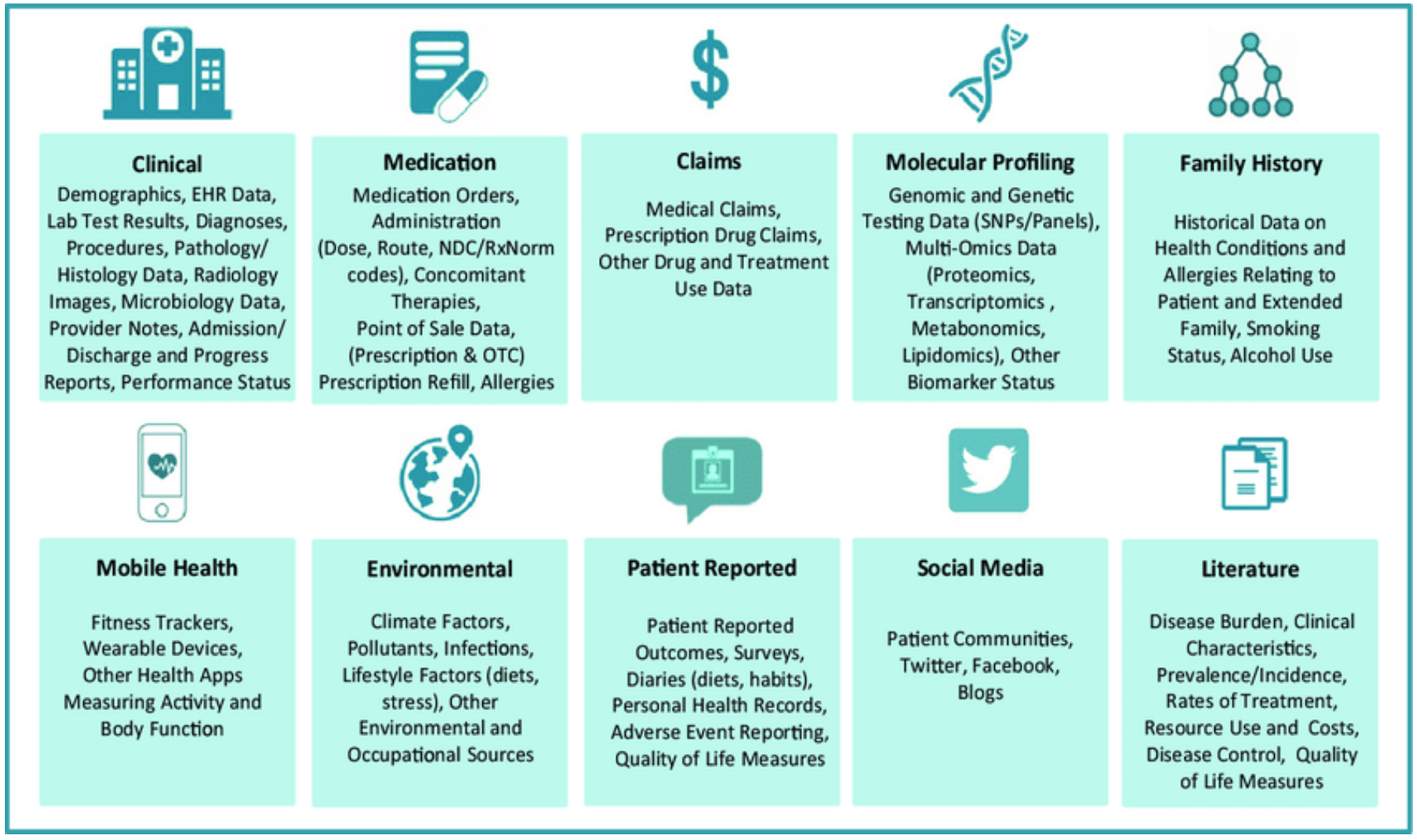
Randomized Controlled Trials (RCTs)



- **RCT** is the **gold standard** for **generating evidence, or answering causal questions,** for medical decision-making,
  - Causal questions: What is the effect of exposure/treatment T on the outcome Y? → Clinical/public health decisions
- However:
  - **Unethical** – Smoking causes lung cancer? Post-acute sequelae of SARS-CoV-2 (Long COVID)? Where did that evidence come from?
  - **$$$** -- 12 million for conducting an RCT on average in drug development
  - **Untimely** -- Studying long-term outcomes takes a long time. E.g., AD progression, long-term post-market efficacy, safety, and adverse events

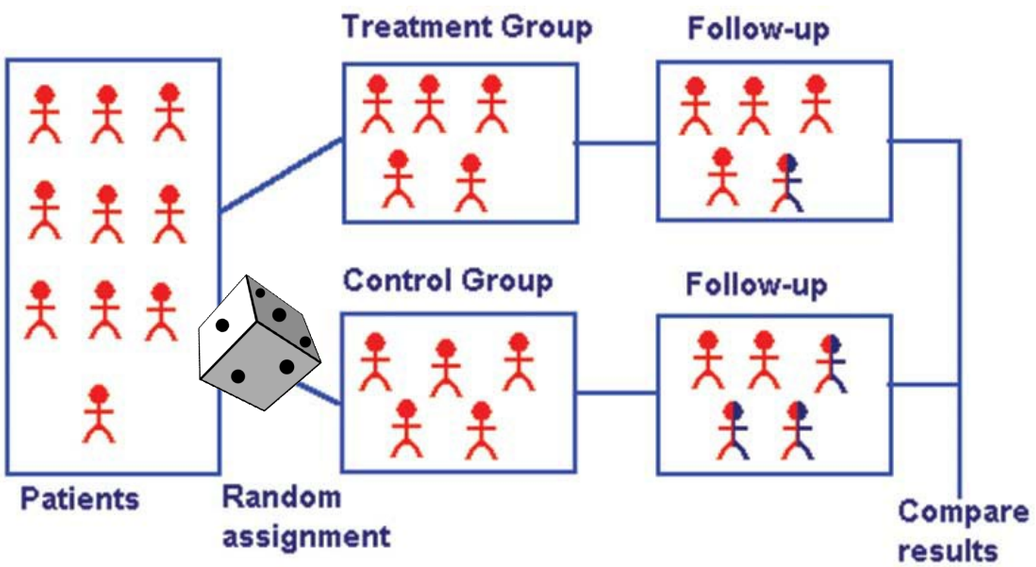
Real-World Data (RWD)
Real-World Evidence (RWE)

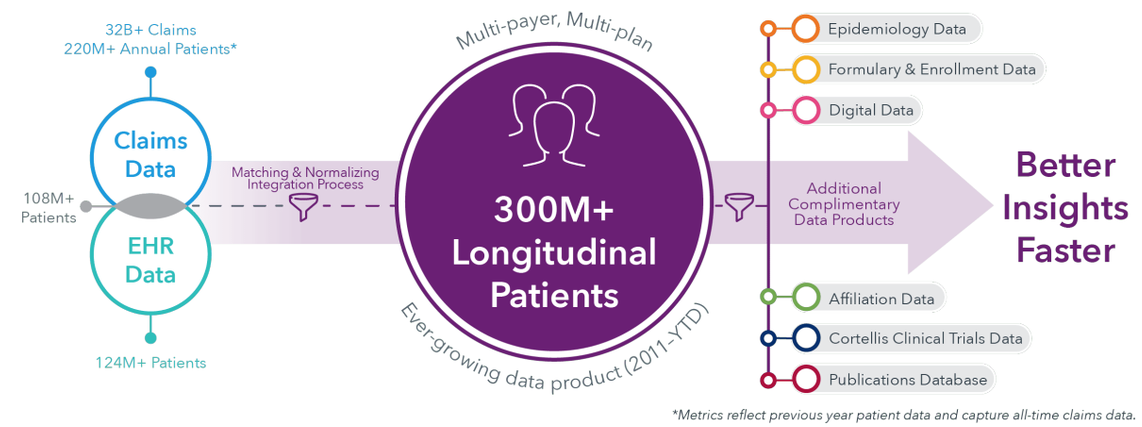
- Real-World Data (RWD)
  - Data relating to patient health status and/or the delivery of health care routinely collected from a variety of sources (patient-level data not collected in conventional RCTs)
    - E.g., EHR, Claims, etc.
- Real-World Evidence (RWE)
  - the clinical evidence derived from the analysis of RWD



FDA 2023
B Swift, CTS 2018

Why Generate RWE using RWD?





- RCT is the gold standard for evidence generation in medical decision-making, however,
- **Unethical** – Smoking causes lung cancer? Post-acute sequelae of SARS-CoV-2 (Long COVID)? Where did that evidence come from?
- **$$$** -- 12 million for conducting an RCT in average
- **Untimely** -- Studying long-term outcomes takes a long time. E.g. AD progression, long-term post-market efficacy, safety and adverse events

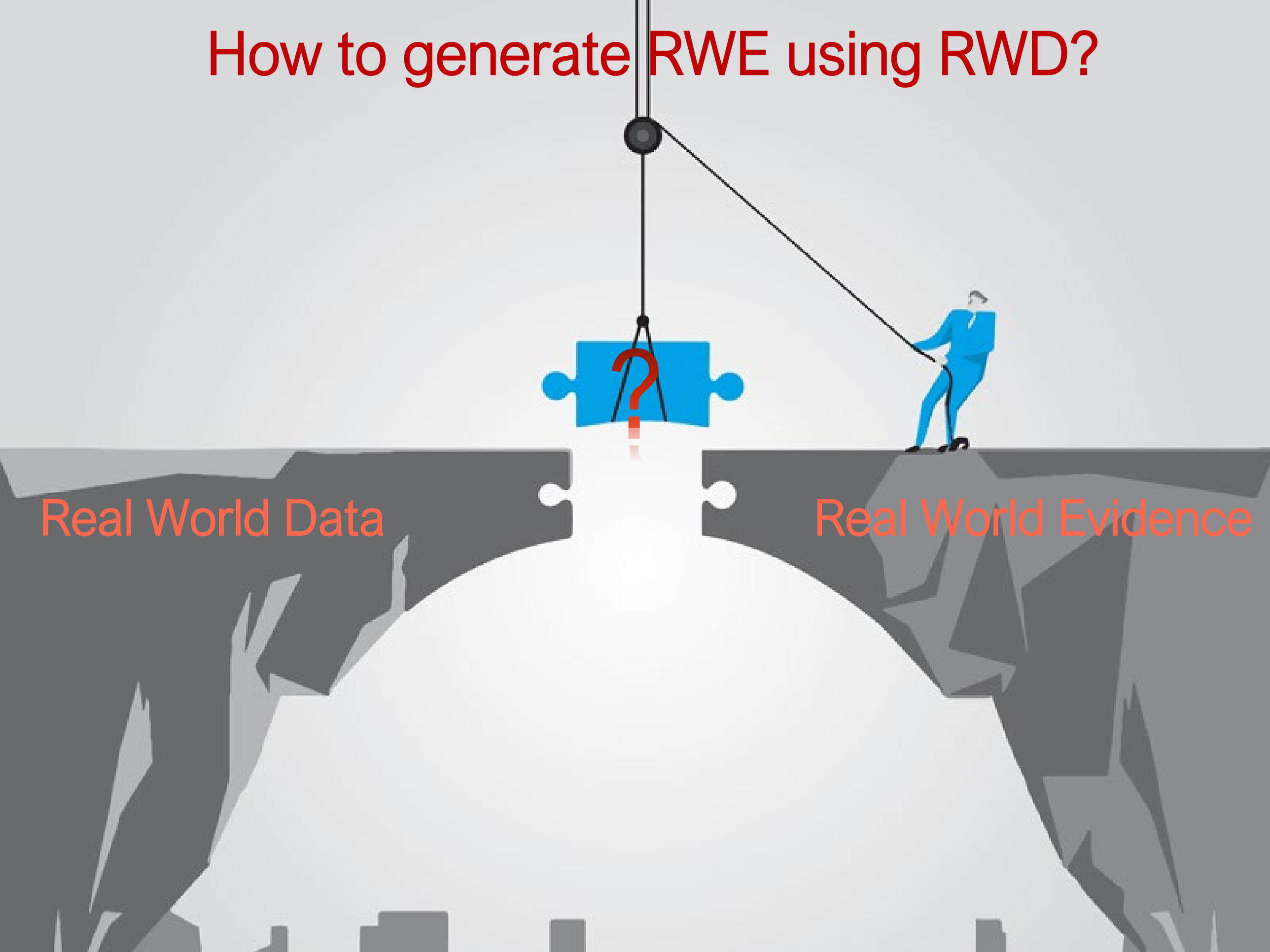
- RWD/RWE → to complement the knowledge gained from traditional clinical trials
  - Observational data, Ethical
  - Timely and long-term
  - Big patient data, generalizability
  - Increase throughput vs. case by case
  - Rare outcomes
- Challenges: Observational study → *Quality,* Non-randomized, *all kinds of biases, missing, censoring, longitudinal, data complexity, etc.*

Why generate RWE using RWD?

	Non-Randomized	Randomized
Non-Interventional	**Observational Study** *Cohort study, case-control study, case-crossover study, etc.* *Define disease, incidence/prevalence, surveillance, risk factors, burden, etc.*	
Interventional	**Externally controlled trial** *Single-group trial with external control group derived from RWD* **Trial emulation** *Drug RWE (e.g., effectiveness in the long term, general population, e.g., covid vaccine), drug repurposing, comparative effectiveness, Post-market safety, effectiveness monitoring*	**RCTs using RWD** *RWD is used to assess enrollment criteria, trial feasibility, recruitment, selection of sites, outcome identification, conduct RCT*

Adapted from NEJM 386.18 (2022): 1680-1682.

How to generate RWE using RWD?



How to generate RWE using RWD? Trial Emulation

- Bridging RWD and RWE (causal answers)
- Dr. Miguel Hernan's Idea:
  - Any causal question can be answered by a randomized trial.
  - Impossible because some RCTs are expensive, Untimely, Unethical, impractical.
  - However, we can do a thought experiment: imagine a hypothetical randomized trial that we would prefer to conduct and analyze, namely, the target trial → emulate this (hypothetical) randomized trial based on RWD

Hernán, M. A., & Robins, J. M. (2016). Using big data to emulate a target trial when a randomized trial is not available. *American journal of epidemiology*, *183*(8), 758-764.

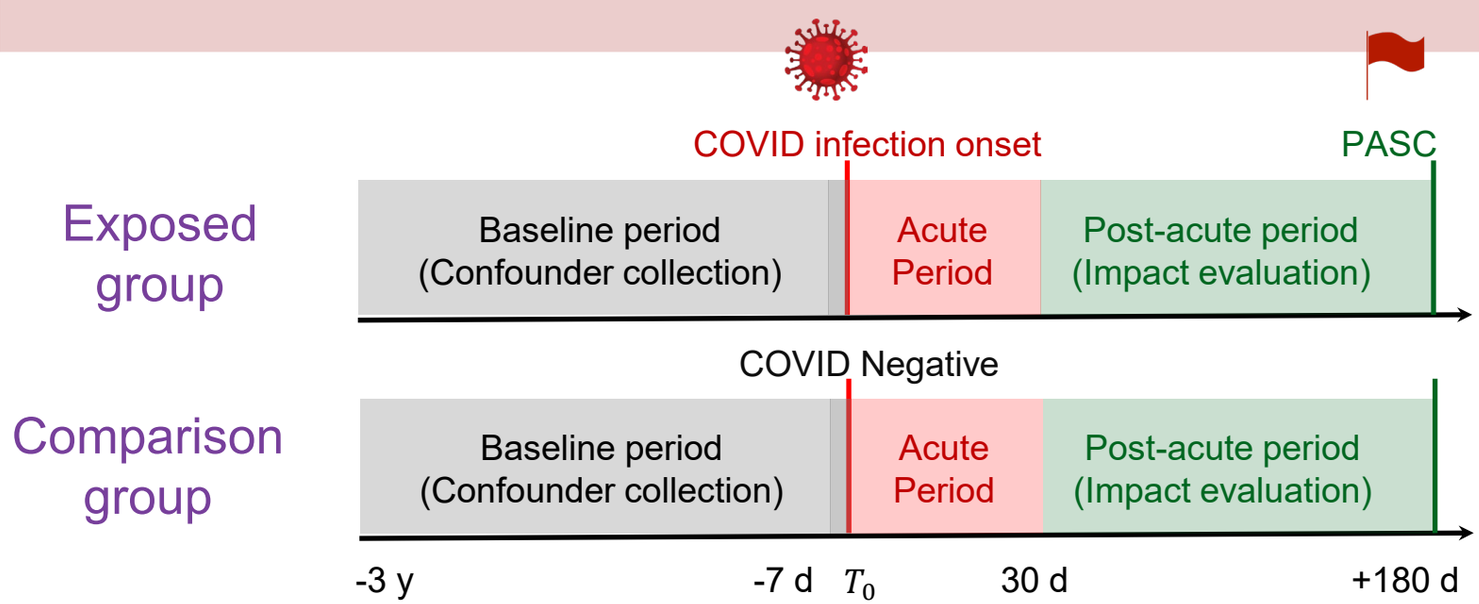
How to generate RWE using RWD? Decompose Trial Emulation

- Step 1: Ask the right causal question(s)
  - Will SARS-CoV-2 infection T lead to incident condition Y in their post-acute period?
- Step 2: Answer that causal question
  - Experiment/Trial design (e.g., https://clinicaltrials.gov/ct2/show/study/NCT04510194)
  - Emulating trial with RWD (Informatics, relevant data sources, feature engineering/ define & validate study variables, faithfully emulated)
  - Causal inference (in lieu of randomization)

Hernán, M. A., & Robins, J. M. (2016). Using big data to emulate a target trial when a randomized trial is not available. *American journal of epidemiology*, *183*(8), 758-764.

Experiment Design: Protocol of a (hypothetical) target trial

Causal questions	Will a SARS-CoV-2 infection (T) lead to incident condition Y (unknown) in their post-acute period?
Eligibility criteria	Adult patients (≥ 20) without lab-confirmed SARS-CoV-2 infection and no history of condition Y in the last 3 years
Exposure strategies	•Exposure group: Infection of SARS-CoV-2 and the SARS-CoV-2 PCR/Antigen tested positive •Control group: No infection of SARS-CoV-2, and the SARS-CoV-2 PCR/Antigen tests kept negative
Assignment	Individuals are randomly assigned to an exposure strategy at baseline and are aware of the assigned exposure strategy.
Follow-up	We followed each patient from his/her infection until the day of the outcome of interest, death, 180 days after baseline, whichever happens first.
Outcomes	Newly-onset post-acute sequelae of COVID-19 (Y).
Causal contrasts	The PASC outcomes were ascertained from day 30 after the SARS-CoV-2 infection and all the causal contrast measures were computed 180 days after the SARS-CoV-2 infection against control group.
Data analysis	Cumulative incidence, excess burden, adjusted hazard ratio, subgroup analyses, sensitivity analysis



Wait! What is Y?

- Goal is to study: how T (covid infection) leads to Y (Long COVID)
- What are Y_s?
  - Build an exhaustive list of all potential PASC conditions
  - Trial emulation for each Ys
  - Prioritize most likely a set of Ys to characterize PASC

Screening List of dx and meds

Abdominal pain and other digestive/abdomen signs and symptoms	Coma; stupor; and brain damage	Intestinal obstruction and ileus	Other specified and unspecified disorders of stomach and duodenum	Retinal and vitreous conditions
Abnormal findings related to substance use	Coronary atherosclerosis and other heart disease	Malaise and fatigue	Other specified and unspecified disorders of the ear	Schizophrenia spectrum and other psychotic disorders
Acquired deformities (excluding foot)	Crystal arthropathies (excluding gout)	Malnutrition	Other specified and unspecified gastrointestinal disorders	Sedative-related disorders
Acquired foot deformities	Depressive disorders	Mediastinal disorders	Other specified and unspecified liver disease	Sequela of specified nervous system conditions
Acute and chronic tonsillitis	Diabetes mellitus with complication	Miscellaneous mental and behavioral disorders/conditions	Other specified and unspecified lower respiratory disease	Sinusitis
Acute and unspecified renal failure	Diabetes mellitus without complication	Muscle disorders	Other specified and unspecified mood disorders	Skin and subcutaneous tissue infections
Acute bronchitis	Diseases of inner ear and related conditions	Musculoskeletal pain, not low back pain	Other specified and unspecified skin disorders	Skin/Subcutaneous signs and symptoms
Acute myocardial infarction	Diseases of the Genitourinary System	Myocarditis and cardiomyopathy	Other specified bone disease and musculoskeletal deformities	Sleep wake disorders
Acute phlebitis; thrombophlebitis and thromboembolism	Drug induced or toxic related condition	Myopathies	Other specified connective tissue disease	Spondylopathies/spondyloarthropathy (including infective)
Acute pulmonary embolism	Encephalitis	Nausea and vomiting	Other specified inflammatory condition of skin	Stimulant-related disorders
Alcohol-related disorders	Epilepsy; convulsions	Nephritis; nephrosis; renal sclerosis	Other specified joint disorders	Symptoms of mental and substance use conditions
Allergic reactions	Esophageal disorders	Nerve and nerve root disorders	Other specified upper respiratory infections	Syncope
Allergic reactions, subsequent encounter	Exposure, encounters, screening or contact with infectious disease	Nervous system pain and pain syndromes	Other substance abuse	Tendon and synovial disorders
Anemia	Feeding and eating disorders	Nervous system signs and symptoms	Otitis media	Tobacco-related disorders
Anxiety and fear-related disorders	Fever	Neurocognitive disorders	Pancreatic disorders (excluding diabetes)	Toxic effects, subsequent encounter
Aortic and peripheral arterial embolism or thrombosis	Fluid and electrolyte disorders	Neurodevelopmental disorders	Paralysis (other than cerebral palsy)	Transient cerebral ischemia
Arterial dissections	Gangrene	Noninfectious hepatitis	PASC-General	Trauma- and stressor-related disorders
Aseptic necrosis and osteonecrosis	Gastritis and duodenitis	Nonspecific chest pain	Pericarditis and pericardial disease	Urinary incontinence
Aspiration pneumonitis	Gastroduodenal ulcer	Obsessive-compulsive and related disorders	Peripheral and visceral vascular disease	Urinary tract infections
Asthma	General sensation/perception signs and symptoms	Occlusion or stenosis of precerebral or cerebral arteries without infarction	Peritonitis and intra-abdominal abscess	Vasculitis
Autoinflammatory syndromes	Genitourinary signs and symptoms	Opioid-related disorders	Pleurisy, pleural effusion and pulmonary collapse	Viral infection
Biliary tract disease	Gout	Osteoarthritis	Pneumonia (except that caused by tuberculosis)	
Bipolar and related disorders	Hallucinogen-related disorders	Other and ill-defined cerebrovascular disease	Pneumothorax	
Cannabis-related disorders	Headache; including migraine	Other and ill-defined heart disease	Polyneuropathies	
Cardiac arrest and ventricular fibrillation	Hearing loss	Other general signs and symptoms	Postthrombotic syndrome and venous insufficiency/hypertension	
Cardiac dysrhythmias	Heart failure	Other nervous system disorders (neither hereditary nor degenerative)	Pressure ulcer of skin	
Cerebral infarction	Hepatic failure	Other nervous system disorders (often hereditary or degenerative)	Pulmonary heart disease	
Chronic obstructive pulmonary disease and bronchiectasis	Hypotension	Other specified and unspecified circulatory disease	Respiratory failure; insufficiency; arrest	
Circulatory signs and symptoms	Immune-mediated/reactive arthropathies	Other specified and unspecified diseases of kidney and ureters	Respiratory signs and symptoms	

vitamin C	General
guaifenesin	Diseases of the Respiratory System
benzonatate	Diseases of the Respiratory System
dextromethorphan	Diseases of the Respiratory System
insulin glargine	Endocrine, Nutritional and Metabolic Diseases
ibuprofen	General
fluticasone	Diseases of the Respiratory System
azithromycin	General
apixaban	Diseases of the Circulatory System
albuterol	Diseases of the Respiratory System
LMW Heparin	Diseases of the Circulatory System
simethicone	Diseases of the Digestive System
laxative	Diseases of the Digestive System
enoxaparin	Diseases of the Circulatory System
insulin lispro	Endocrine, Nutritional and Metabolic Diseases
melatonin	Diseases of the Nervous System
acetaminophen	General
glucagon	Endocrine, Nutritional and Metabolic Diseases
sennosides	Diseases of the Digestive System
witch hazel	Diseases of the Skin and Subcutaneous Tissue
metformin	Endocrine, Nutritional and Metabolic Diseases
ferrous cation	Diseases of the Circulatory System
collagenase	Diseases of the Skin and Subcutaneous Tissue
budesonide	Diseases of the Respiratory System
vitamin D3	General
formoterol	Diseases of the Respiratory System
vilanterol trifenatate	Diseases of the Respiratory System
ipratropium	Diseases of the Respiratory System
prednisone	Diseases of the Respiratory System
aluminum hydroxide	Diseases of the Digestive System
magnesium hydroxide	Diseases of the Digestive System

CLINICAL CLASSIFICATIONS SOFTWARE REFINED (CCSR) FOR ICD-10-CM DIAGNOSES, v2022.1 with 73,371 ICD-10-CM does, 530 categories → selected by the clinician group → 6,000+ ICD-10-CM codes with 137 categories for adults

459 top prevalent medications at rxNorm at ingredient level from INSIGHT & OneFlorida+

Increase Throughput

Y_{596}

Eligibility criteria	Adult patients (≥ 20) with lab-confirmed SARS-CoV-2 infection and no history of condition Y in the last 3 years
Exposure strategies	•Exposure group: Infection of SARS-CoV-2 and the SARS-CoV-2 PCR/Antigen tested positive •Control group: No infection of SARS-CoV-2, and the SARS-CoV-2 PCR/Antigen tests kept negative
Assignment	Individuals are randomly assigned to an exposure strategy at baseline and are aware of the assigned exposure strategy.
Follow-up	We followed each patient from his/her baseline day until the day of the outcome of interest, death, 180 days after baseline, or the end of the study period (November 30, 2021), whichever happens first.

Y_2

Eligibility criteria	Adult patients (≥ 20) with lab-confirmed SARS-CoV-2 infection and no history of condition Y in the last 3 years
Exposure strategies	•Exposure group: Infection of SARS-CoV-2 and the SARS-CoV-2 PCR/Antigen tested positive •Control group: No infection of SARS-CoV-2, and the SARS-CoV-2

Y_1

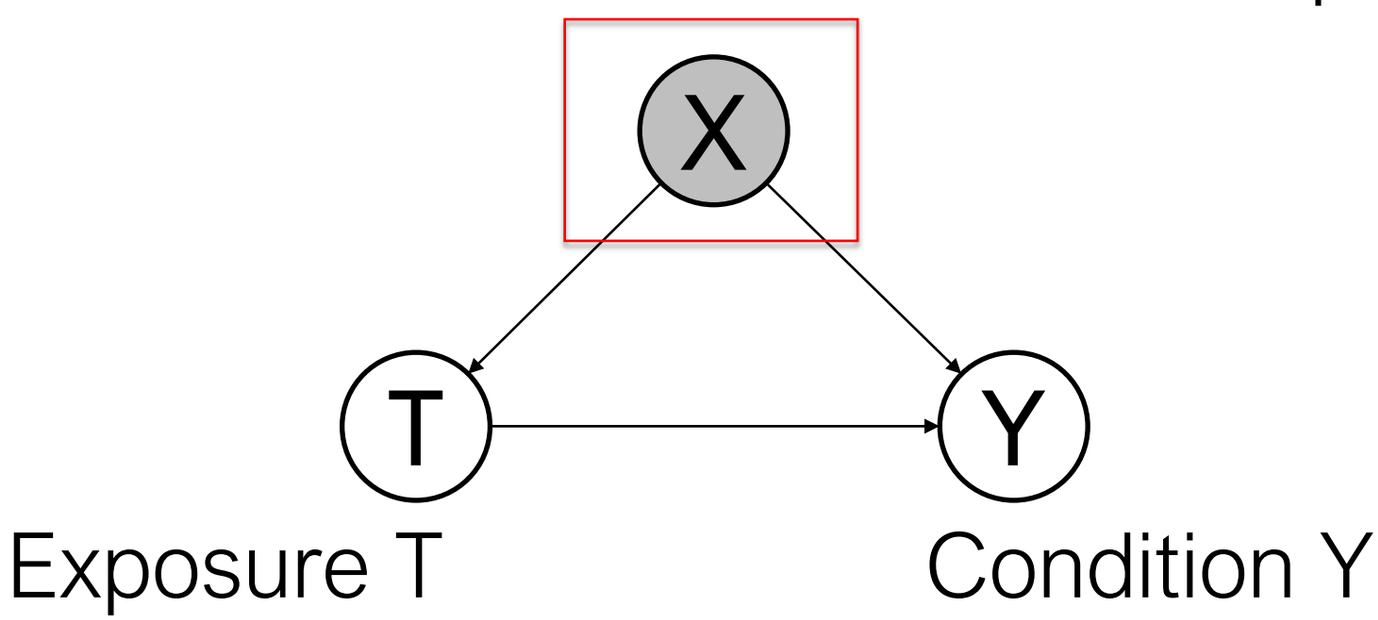
Eligibility criteria	Adult patients (≥ 20) with lab-confirmed SARS-CoV-2 infection and no history of condition Y in the last 3 years
Exposure strategies	•Exposure group: Infection of SARS-CoV-2 and the SARS-CoV-2 PCR/Antigen tested positive •Control group: No infection of SARS-CoV-2, and the SARS-CoV-2 PCR/Antigen tests kept negative
Assignment	Individuals are randomly assigned to an exposure strategy at baseline and are aware of the assigned exposure strategy.
Follow-up	We followed each patient from his/her baseline day until the day of the outcome of interest, death, 180 days after baseline, or the end of the study period (November 30, 2021), whichever happens first.
Outcomes	Newly-onset post-acute sequelae of COVID-19 (Y).
Causal contrasts	The PASC outcomes were ascertained from day 30 after the SARS-CoV-2 infection and all the adjusted risk measures were computed 180 days after the SARS-CoV-2 infection against control group.
Data analysis	Cumulative incidence, excess burden, adjusted hazard ratio, subgroup analyses, sensitivity analysis

137+459 = 596 trials

Wait! Random exposure assignment? Adjust Analysis/Causal Inference

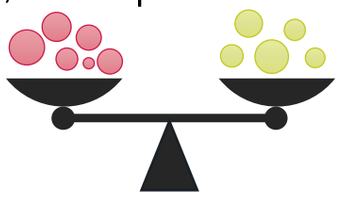
- **Data ($\mathbf{X}, \mathbf{T}, \mathbf{Y}$):** X: baseline covariates; $T \in \{0, 1\}$ treatment/exposure assignment; Y: outcome

X Confounding factor, e.g., baseline conditions, age, gender, race, BMI, social-eco status, index period, etc.



Causal Inference with PS

- **Exposure groups were exchangeable by adjusting for baseline covariates.**
  - Data (X, T, Y), X: baseline covariates including: basic demographics age, race, gender, medications, diagnoses, SDOH etc.; $T \in \{0, 1\}$ treatment assignment; Y: outcome
  - **Identifying Assumptions: conditional exchangeability, positivity, consistency, non-interference**
- **Propensity Score (PS)** $\mathrm{P(T = 1|X)}$: "the conditional probability of assignment to a particular treatment given a vector of observed covariates."
- **Inverse Probability of Treatment Weight (IPTW)** as **sample weights** for adjustment
  - PTW $w = \frac{T}{P} + \frac{1-T}{1-P}$ → Stabilized-IPTW $\mathrm{w} = \frac{T*\mathrm{P(T=1)}}{P} + \frac{(1-T)*\mathrm{P(T=0)}}{1-P}$ → clipped 0.01,0.99 quantiles
  - Patients re-weighted by w → a pseudo-Randomized Controlled Trial
  - Adjusted outcome, e.g., hazard ratio, excess burden, etc.



- Balance diagnostics: Standardized Mean Difference
- **AI/ML/DL:**
  - Learning PS is a binary classification problem
  - $P_\Theta : X \rightarrow T$ with learnable parameter Θ
  - Naïve idea: Can we propose more powerful/complex/deep P_Θ ?

Outline

1. Background

2. Key Concepts & Causal Basics

3. **Applications & Beyond**

   - **To Characterize Long COVID in terms of individual and clustered conditions**

4. Conclusions

To define Long COVID through Data-driven High-throughput Analysis

pcornet® The National Patient-Centered Clinical Research Network

RECOVER Researching COVID to Enhance Recovery
An Initiative Funded by the National Institutes of Health

nature communications

Article

https://doi.org/10.1038/s41467-023-37653-z

Data-driven analysis to understand long COVID using electronic health records from the RECOVER initiative

Received: 8 June 2022

Accepted: 24 March 2023

Published online: 07 April 2023

Check for updates

Chengxi Zang[1], Yongkang Zhang[1], Jie Xu[2], Jiang Bian[2], Dmitry Morozyuk[1], Edward J. Schenck[3], Dhruv Khullar[1], Anna S. Nordvig[4], Elizabeth A. Shenkman[2], Russell L. Rothman[5], Jason P. Block[6], Kristin Lyman[7], Mark G. Weiner[1], Thomas W. Carton[7], Fei Wang[1] & Rainu Kaushal[1]

Recent studies have investigated post-acute sequelae of SARS-CoV-2 infection (PASC, or long COVID) using real-world patient data such as electronic health records (EHR). Prior studies have typically been conducted on patient cohorts with specific patient populations which makes their generalizability unclear. This study aims to characterize PASC using the EHR data warehouses from two large Patient-Centered Clinical Research Networks (PCORnet), INSIGHT and OneFlorida+, which include 11 million patients in New York City (NYC) area and 16.8 million patients in Florida respectively. With a high-throughput screening pipeline based on propensity score and inverse probability of treatment weighting, we identified a broad list of diagnoses and medications which exhibited significantly higher incidence risk for patients 30–180 days after the laboratory-confirmed SARS-CoV-2 infection compared to non-infected patients. We identified more PASC diagnoses in NYC than in Florida regarding our screening criteria, and conditions including dementia, hair loss, pressure ulcers, pulmonary fibrosis, dyspnea, pulmonary embolism, chest pain, abnormal heartbeat, malaise, and fatigue, were replicated across both cohorts. Our analyses highlight potentially heterogeneous risks of PASC in different populations.

The global COVID-19 pandemic from late 2019 has led to more than 620 million infections and 6.5 million deaths as of Oct 17, 2022[1]. Growing scientific and clinical evidence has demonstrated potential post-acute and long-term effects of SARS-CoV-2 infection in multiple organ systems[2], including cardiovascular[3], mental health[4], neurological[5], and metabolic[6] among other systems. Recently, several retrospective observational cohort analyses have described post-acute sequelae of SARS-CoV-2 infection (PASC) using real-world patient data[7–9]. These studies typically start with a predefined list of PASC symptoms and signs and then contrast their incidence risk or burden in SARS-CoV-2 infected patients versus non-infected controls. Different analytical pipelines have been utilized, such as causal inference[7], regression analysis[10], and network analysis[11]. There are two major challenges to these existing studies. First, the disease etiology and

[1]Department of Population Health Sciences, Weill Cornell Medicine, New York, NY, USA. [2]Department of Health Outcomes Biomedical Informatics, University of Florida, Gainesville, FL, USA. [3]Department of Medicine, Division of Pulmonary and Critical Care Medicine, Weill Cornell Medicine, New York, NY, USA. [4]Department of Neurology, Weill Cornell Medicine, New York, NY, USA. [5]Center for Health Services Research, Vanderbilt University Medical Center, Nashville, TN, USA. [6]Department of Population Medicine, Harvard Pilgrim Health Care Institute, Harvard Medical School, Boston, MA, USA. [7]Louisiana Public Health Institute, New Orleans, LA, USA. e-mail: few2001@med.cornell.edu

To define Long COVID through Data-driven High-throughput Analysis

pcornet
The National Patient-Centered Clinical Research Network

RECOVER
Researching COVID to Enhance Recovery
An Initiative Funded by the National Institutes of Health

nature COMMUNICATIONS

Objectives

- Characterize Long COVID through increased risk of new EHR diagnoses and medications in a SARS-CoV-2 patients compared with controls in NYC and Florida

Methods

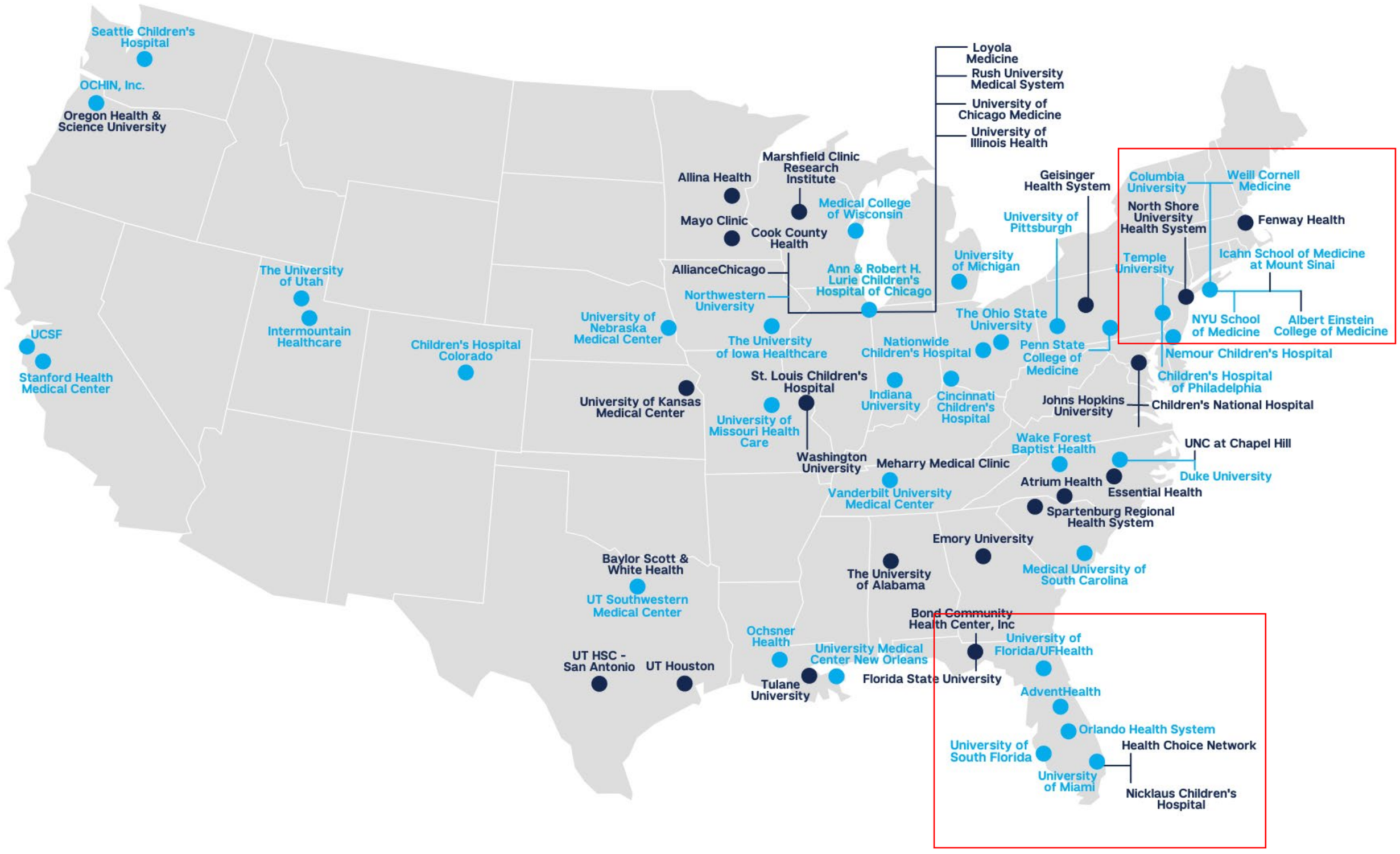
- Created a list of 137 PASC associated diagnoses and 459 medications based on literature review and expert clinical consultation
- High-throughput causal inference pipeline using high-dimensional inverse propensity score adjustment to compare the new incidence of these codes in 57,616 SARS-CoV-2 infected patients from 31 days to 180 days after their acute infection compared with 503,136 controls
  - Included patients with at least one SARS-CoV-2 polymerase-chain-reaction (PCR) or antigen laboratory test between March 01, 2020, and November 30, 2021
- Studied a large population in New York City (14m) and Florida (17m)

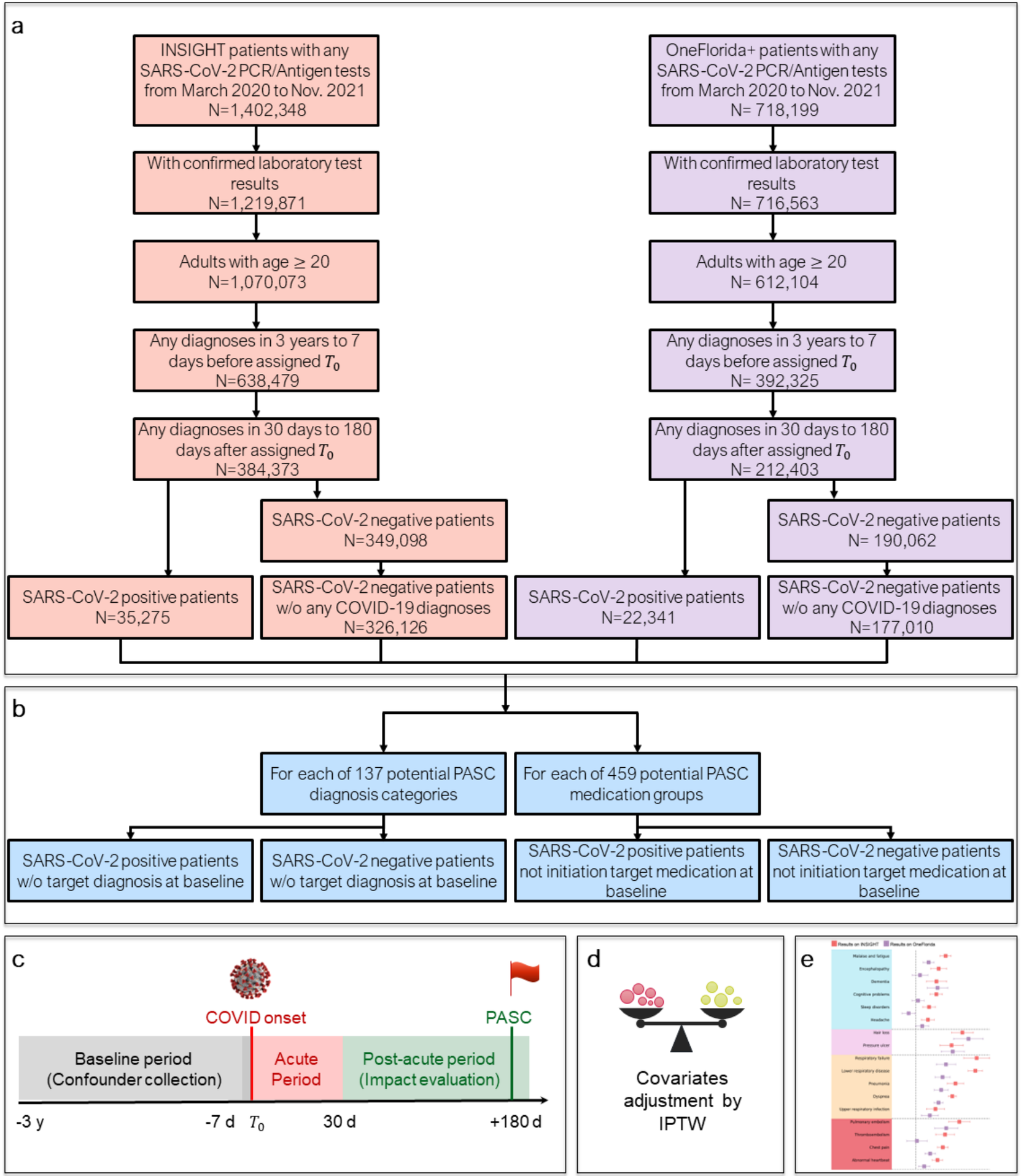
"Data-driven analysis to understand long COVID using electronic health records from the RECOVER initiative." *Nature Communications* 14.1 (2023): 1948.

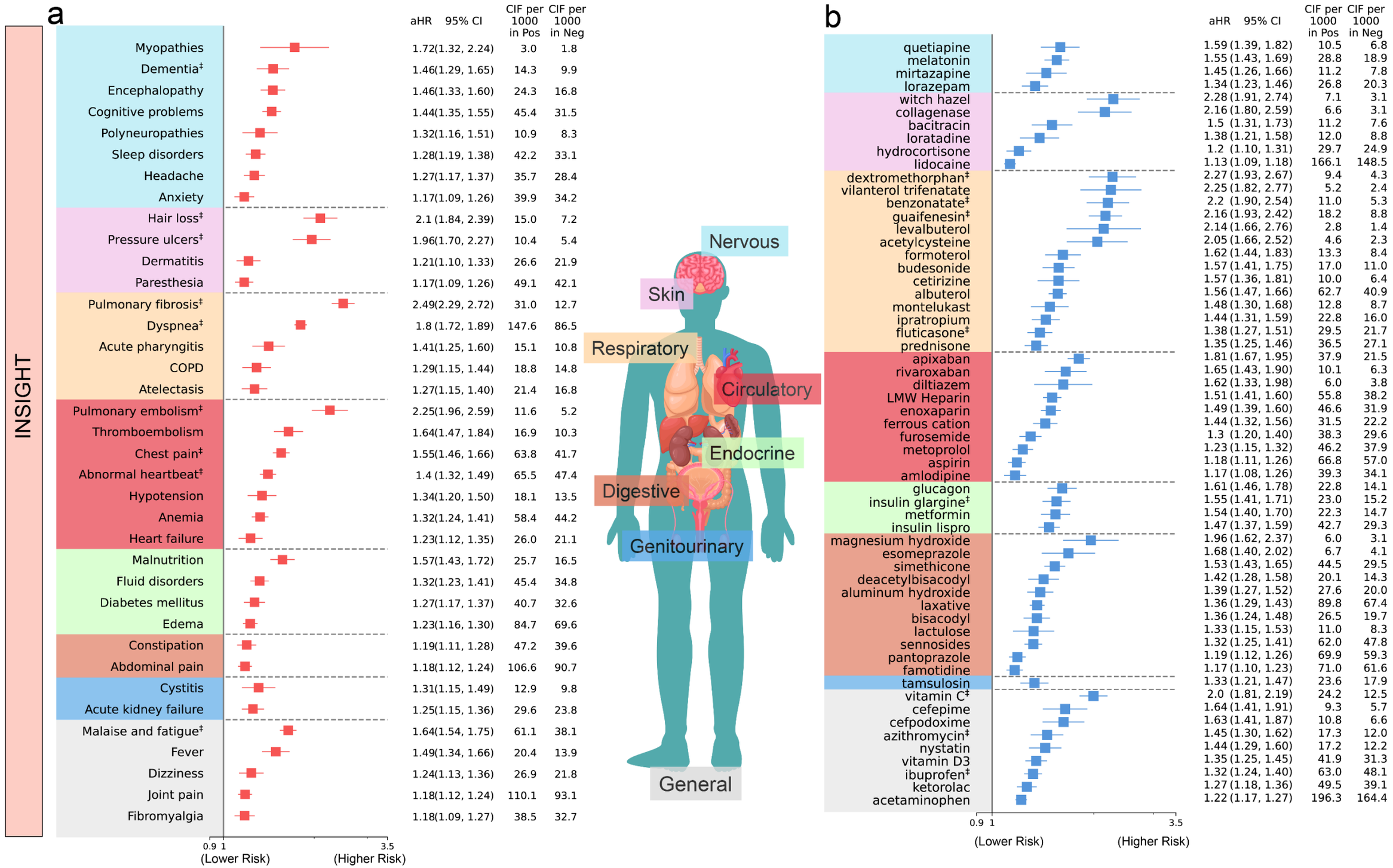
Leveraging EHR/RWD to Understand Long COVID

pcornet®
The National Patient-Centered Clinical Research Network

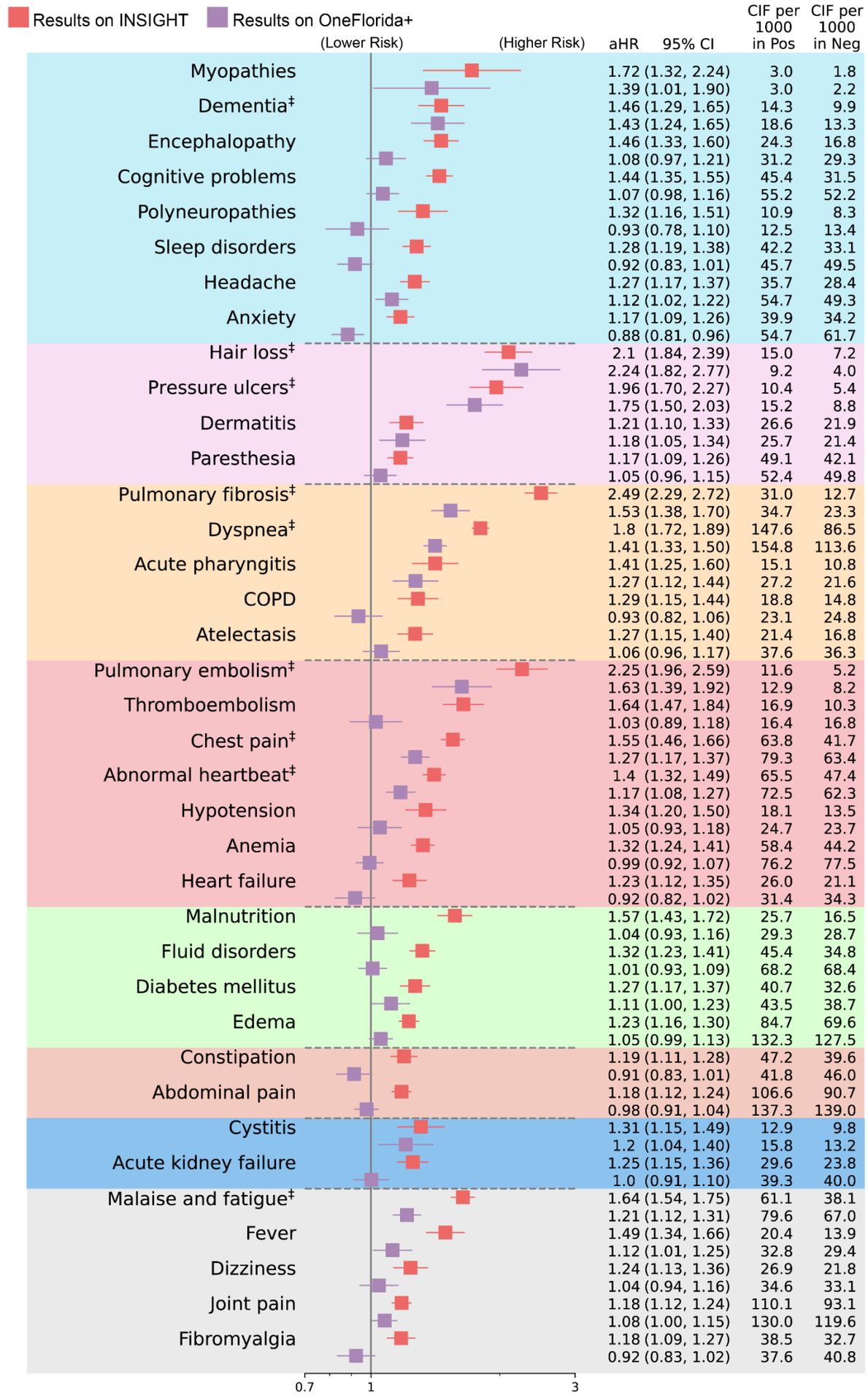
RECOVER
Researching COVID to Enhance Recovery
An Initiative Funded by the National Institutes of Health




a
INSIGHT patients with any SARS-CoV-2 PCR/Antigen tests from March 2020 to Nov. 2021 N=1,402,348
With confirmed laboratory test results N=1,219,871
Adults with age ≥ 20 N=1,070,073
Any diagnoses in 3 years to 7 days before assigned T_0 N=638,479
Any diagnoses in 30 days to 180 days after assigned T_0 N=384,373
SARS-CoV-2 negative patients N=349,098
SARS-CoV-2 positive patients N=35,275
SARS-CoV-2 negative patients w/o any COVID-19 diagnoses N=326,126
OneFlorida+ patients with any SARS-CoV-2 PCR/Antigen tests from March 2020 to Nov. 2021 N= 718,199
With confirmed laboratory test results N= 716,563
Adults with age ≥ 20 N= 612,104
Any diagnoses in 3 years to 7 days before assigned T_0 N= 392,325
Any diagnoses in 30 days to 180 days after assigned T_0 N= 212,403
SARS-CoV-2 negative patients N= 190,062
SARS-CoV-2 positive patients N=22,341
SARS-CoV-2 negative patients w/o any COVID-19 diagnoses N=177,010
b
For each of 137 potential PASC diagnosis categories
For each of 459 potential PASC medication groups
SARS-CoV-2 positive patients w/o target diagnosis at baseline
SARS-CoV-2 negative patients w/o target diagnosis at baseline
SARS-CoV-2 positive patients not initiation target medication at baseline
SARS-CoV-2 negative patients not initiation target medication at baseline
c
COVID onset
PASC
Baseline period (Confounder collection)
Acute Period
Post-acute period (Impact evaluation)
-3 y
-7 d
T_0
30 d
+180 d
d
Covariates adjustment by IPTW
e

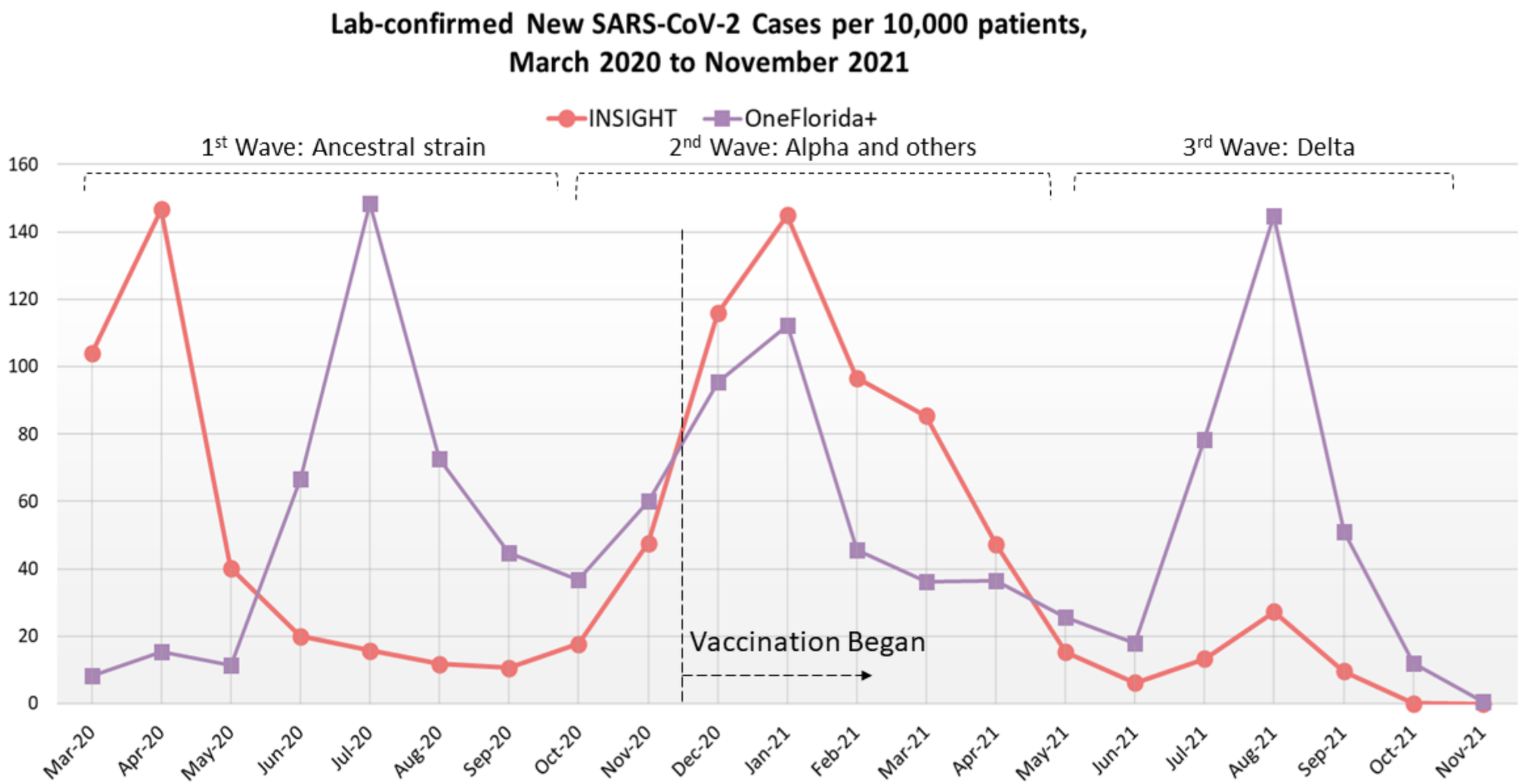


Nature Communication 23



Nature Communication 23

Why different?



	Mar-20	Apr-20	May-20	Jun-20	Jul-20	Aug-20	Sep-20	Oct-20	Nov-20	Dec-20	Jan-21	Feb-21	Mar-21	Apr-21	May-21	Jun-21	Jul-21	Aug-21	Sep-21	Oct-21	Nov-21
INSIGHT	103.9	146.7	40.2	20.0	15.8	11.7	10.6	17.8	47.6	115.9	144.9	96.6	85.4	47.2	15.4	6.2	13.3	27.2	9.6	0.1	0.0
OneFlorida+	8.4	15.5	11.3	66.7	148.5	72.6	44.7	36.9	60.2	95.5	112.4	45.7	36.1	36.6	25.5	18.0	78.3	144.7	50.9	11.8	0.5

Nature Communication 23

Why different?

Table 1. Baseline characteristics of the lab-confirmed SARS-CoV-2 positive patients and SARS-CoV-2 negative patients in the INSIGHT and OneFlorida+ cohorts, March 2020 to November 2021[a].

	INSIGHT			OneFlorida+		
Characteristics	SARS-CoV-2 Positive (N=35,275)	SARS-CoV-2 Negative (N=326,126)	SMD[b]	SARS-CoV-2 Positive (N=22,341)	SARS-CoV-2 Negative (N=177,010)	SMD[b]
Median age (IQR) - years	55 (38-68)	57 (40-69)	-0.09	50 (34-64)	57 (40-69)	-0.27
Age group - no. (%)						
20-<40 years	9,529 (27.0)	77,403 (23.7)	0.08	7,506 (33.6)	42,286 (23.9)	0.22
40-<55 years	7,975 (22.6)	70,313 (21.6)	0.03	5,473 (24.5)	37,555 (21.2)	0.08
55-<65 years	6,965 (19.7)	66,361 (20.3)	-0.02	4,036 (18.1)	37,142 (21.0)	-0.07
65-<75 years	5,712 (16.2)	62,860 (19.3)	-0.08	2,929 (13.1)	34,601 (19.5)	-0.17
75+ years	5,094 (14.4)	49,189 (15.1)	-0.02	2,397 (10.7)	25,426 (14.4)	-0.11
Sex - no. (%) Female	20,686 (58.6)	196,730 (60.3)	-0.03	14,004 (62.7)	106,963 (60.4)	0.05
Male	14,586 (41.3)	129,360 (39.7)	0.03	8,335 (37.3)	70,034 (39.6)	-0.05
Race - no. (%) Asian	1,736 (4.9)	17,439 (5.3)	-0.02	275 (1.2)	2,912 (1.6)	-0.03
Black	7,791 (22.1)	62,281 (19.1)	0.07	6,504 (29.1)	35,381 (20.0)	0.21
White	12,233 (34.7)	139,512 (42.8)	-0.17	11,398 (51.0)	105,521 (59.6)	-0.17
Other	9,844 (27.9)	69,406 (21.3)	0.15	3,730 (16.7)	30,138 (17.0)	-0.01
Missing	3,671 (10.4)	37,488 (11.5)	-0.03	434 (1.9)	3,058 (1.7)	0.02
Ethnic group - no. (%)						
Hispanic	10,658 (30.2)	73,522 (22.5)	0.17	4,500 (20.1)	21,484 (12.1)	0.22
Not Hispanic	20,838 (59.1)	216,179 (66.3)	-0.15	14,798 (66.2)	120,315 (68.0)	-0.04
Unknown	3,779 (10.7)	36,425 (11.2)	-0.01	3,043 (13.6)	35,211 (19.9)	-0.17
Median ADI (IQR) - rank	15 (6-24)	13 (5-23)	0.03	58 (41-76)	53 (36-72)	0.19
BMI kg/m² (IQR)	27 (21-32)	25 (1-30)	0.02	30 (25-35)	28 (24-34)	0.00
Follow-up days (IQR)	258 (163-418)	269 (145-388)	0.09	207 (109-367)	250 (122-409)	-0.17
Cares in the past 3 years — no. (%)						
Inpatient 0	25,717 (72.9)	278,784 (85.5)	-0.31	12,838 (57.5)	112,480 (63.5)	-0.12
Inpatient 1-2	6,805 (19.3)	37,297 (11.4)	0.22	4,614 (20.7)	33,658 (19.0)	0.04
Inpatient >=3	2,753 (7.8)	10,045 (3.1)	0.21	4,889 (21.9)	30,872 (17.4)	0.11
Corticosteroids Prescription	4,999 (14.2)	28,915 (8.9)	0.17	4,253 (19.0)	27,783 (15.7)	0.09
Immunosuppressant Prescriptions	2,110 (6.0)	10,761 (3.3)	0.13	1,013 (4.5)	7,281 (4.1)	0.02

Nature Communication 23

To define Long COVID through Data-driven High-throughput Analysis

pcornet
The National Patient-Centered Clinical Research Network

RECOVER
Researching COVID to Enhance Recovery
An Initiative Funded by the National Institutes of Health

nature COMMUNICATIONS

Objectives

- Characterize Long COVID through increased risk of new EHR diagnoses and medications in a SARS-CoV-2 patients compared with controls in NYC and Florida

Methods

- Created a list of 137 PASC associated diagnoses and 459 medications based on literature review and expert clinical consultation
- High-throughput causal inference pipeline using high-dimensional inverse propensity score adjustment to compare the new incidence of these codes in 57,616 SARS-CoV-2 infected patients from 31 days to 180 days after their acute infection compared with 503,136 controls
  - Included patients with at least one SARS-CoV-2 polymerase-chain-reaction (PCR) or antigen laboratory test between March 01, 2020, and November 30, 2021
- Studied a large population in New York City (14m) and Florida (17m)

Results

- Identified significantly higher incidence of conditions in **multiple organ systems**: respiratory, circulatory, musculoskeletal & connective tissue, neurological disorders, psychiatric, gastrointestinal, endocrine, metabolic, blood, genitourinary
- Higher burden of PASC in NYC compared with Florida

"Data-driven analysis to understand long COVID using electronic health records from the RECOVER initiative." *Nature Communications* 14.1 (2023): 1948.

Conquer heterogeneity by Sub-phenotyping Long COVID

pcornet The National Patient-Centered Clinical Research Network

RECOVER Researching COVID to Enhance Recovery

An Initiative Funded by the National Institutes of Health

nature medicine

Article

https://doi.org/10.1038/s41591-022-02116-3

Data-driven identification of post-acute SARS-CoV-2 infection subphenotypes

Received: 8 June 2022

Accepted: 2 November 2022

Published online: 1 December 2022

Check for updates

Hao Zhang[1], Chengxi Zang[1], Zhenxing Xu[1], Yongkang Zhang[1], Jie Xu[2], Jiang Bian[2], Dmitry Morozyuk[1], Dhruv Khullar[1], Yiye Zhang[1], Anna S. Nordvig[3], Edward J. Schenck[4], Elizabeth A. Shenkman[2], Russell L. Rothman[5], Jason P. Block[6], Kristin Lyman[7], Mark G. Weiner[1], Thomas W. Carton[7], Fei Wang[1] & Rainu Kaushal[1]

The post-acute sequelae of SARS-CoV-2 infection (PASC) refers to a broad spectrum of symptoms and signs that are persistent, exacerbated or newly incident in the period after acute SARS-CoV-2 infection. Most studies have examined these conditions individually without providing evidence on co-occurring conditions. In this study, we leveraged the electronic health record data of two large cohorts, INSIGHT and OneFlorida+, from the national Patient-Centered Clinical Research Network. We created a development cohort from INSIGHT and a validation cohort from OneFlorida+ including 20,881 and 13,724 patients, respectively, who were SARS-CoV-2 infected, and we investigated their newly incident diagnoses 30–180 days after a documented SARS-CoV-2 infection. Through machine learning analysis of over 137 symptoms and conditions, we identified four reproducible PASC subphenotypes, dominated by cardiac and renal (including 33.75% and 25.43% of the patients in the development and validation cohorts); respiratory, sleep and anxiety (32.75% and 38.48%); musculoskeletal and nervous system (23.37% and 23.35%); and digestive and respiratory system (10.14% and 12.74%) sequelae. These subphenotypes were associated with distinct patient demographics, underlying conditions before SARS-CoV-2 infection and acute infection phase severity. Our study provides insights into the heterogeneity of PASC and may inform stratified decision-making in the management of PASC conditions.

The ongoing global pandemic of Coronavirus Disease 2019 (COVID-19) caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) infection has impacted hundreds of millions of people's lives. Existing studies have provided evidence that many symptoms and signs could be persistent, exacerbated or newly present after the acute phase of SARS-CoV-2 infection, referred to as post-acute sequelae of SARS-CoV-2 infection (PASC)[1,2], which involve multiple organ systems, including cardiovascular[3], mental[4], metabolic[5], renal[6] and others. There have been various ongoing efforts into investigating the underlying biological mechanisms of PASC[7–9], which have typically been conducted in small patient cohorts. Large-scale clinical observational cohort studies can provide useful insights into PASC that may help develop

[1]Department of Population Health Sciences, Weill Cornell Medicine, New York, NY, USA. [2]Department of Health Outcomes Biomedical Informatics, University of Florida, Gainesville, FL, USA. [3]Department of Neurology, Weill Cornell Medicine, New York, NY, USA. [4]Department of Medicine, Division of

Conquer heterogeneity by Sub-phenotyping Long COVID

pcornet
The National Patient-Centered Clinical Research Network

RECOVER
Researching COVID to Enhance Recovery
An Initiative Funded by the National Institutes of Health

nature medicine

Objectives

- Utilize topic modeling to identify sub-phenotypes of Long COVID

Methods

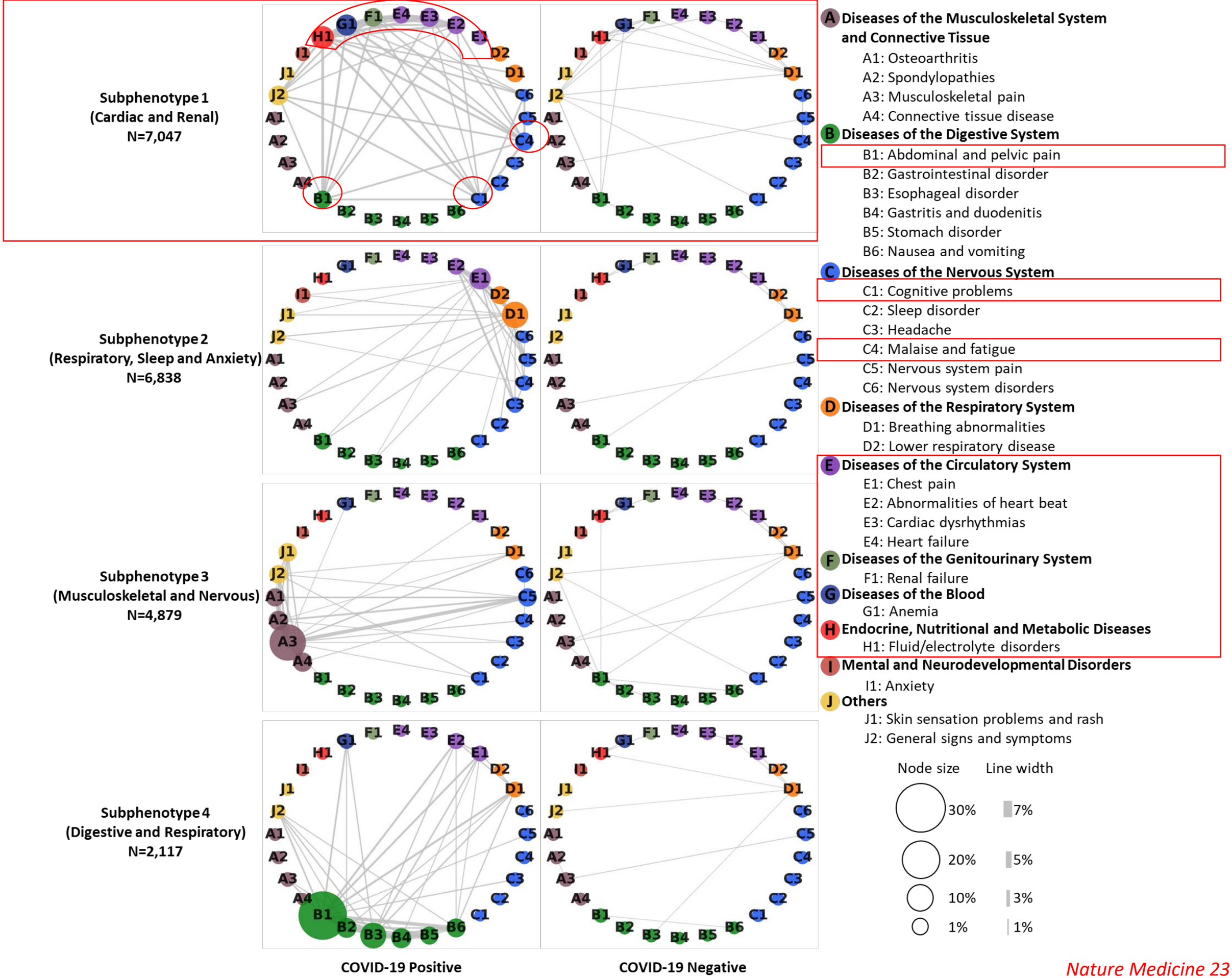
- Mapped 137 newly incident diagnoses into 10 topics based on co-occurrence patterns in 34,605 patients via Topic Modeling (Poisson Factor Analysis)
- Analyzed clustering of topics in patients to demonstrate four sub-phenotypes
- Compared with matched controls (age, gender, race, ADI, exact match, others PS-match)

Results

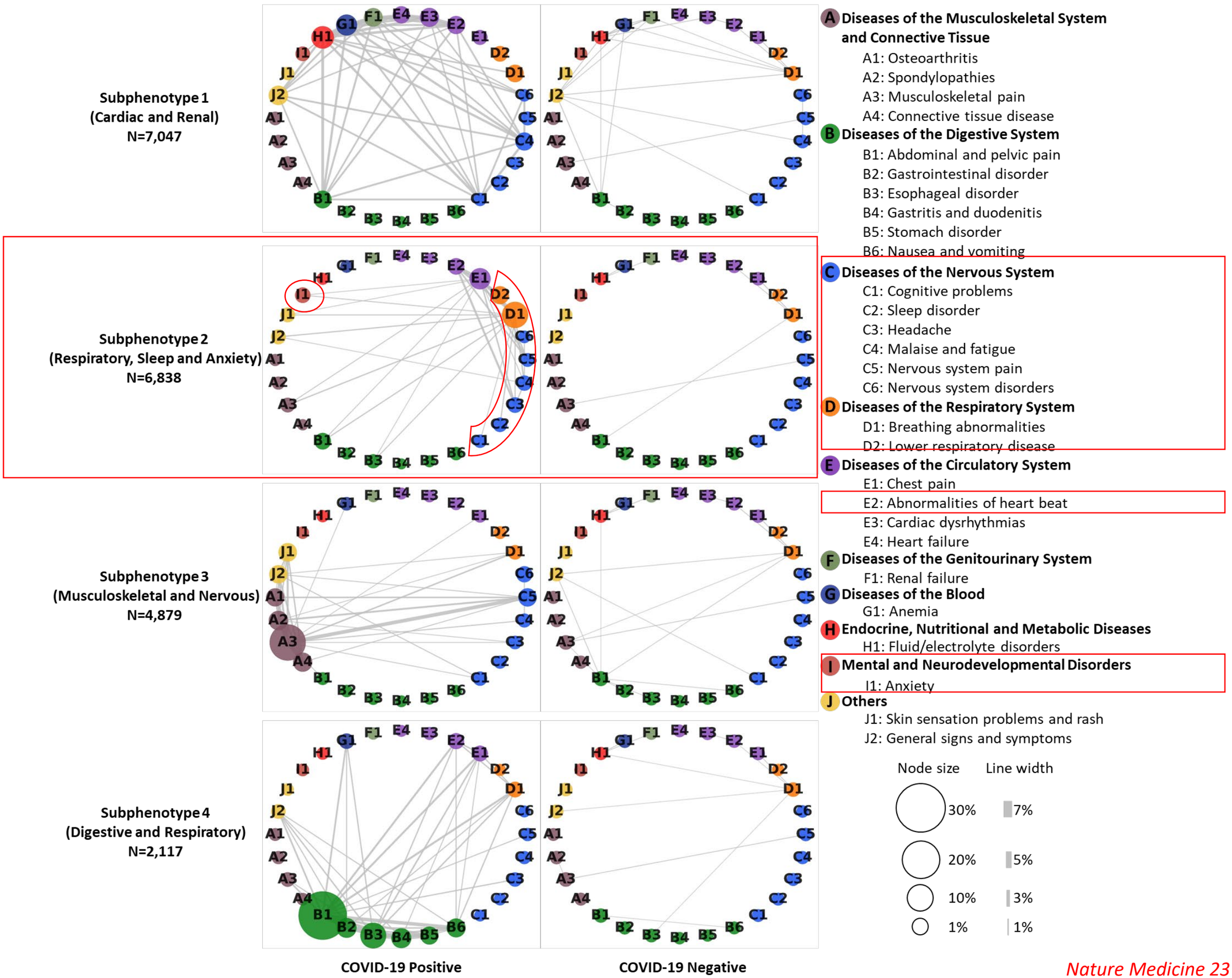
- Four sub-phenotypes characterized by:

1. **Cardiac and renal** (median age 65, 51% female, higher severity in acute phase)
2. **Respiratory conditions, sleep disorders & anxiety** (median age 51, 63% female, lowest rates of hospitalization)
3. **Musculoskeletal and nervous system** (median age 57, 61% female)
4. **Digestive system and respiratory conditions** (median age 54, 62% female, lowest rates of ICU care)

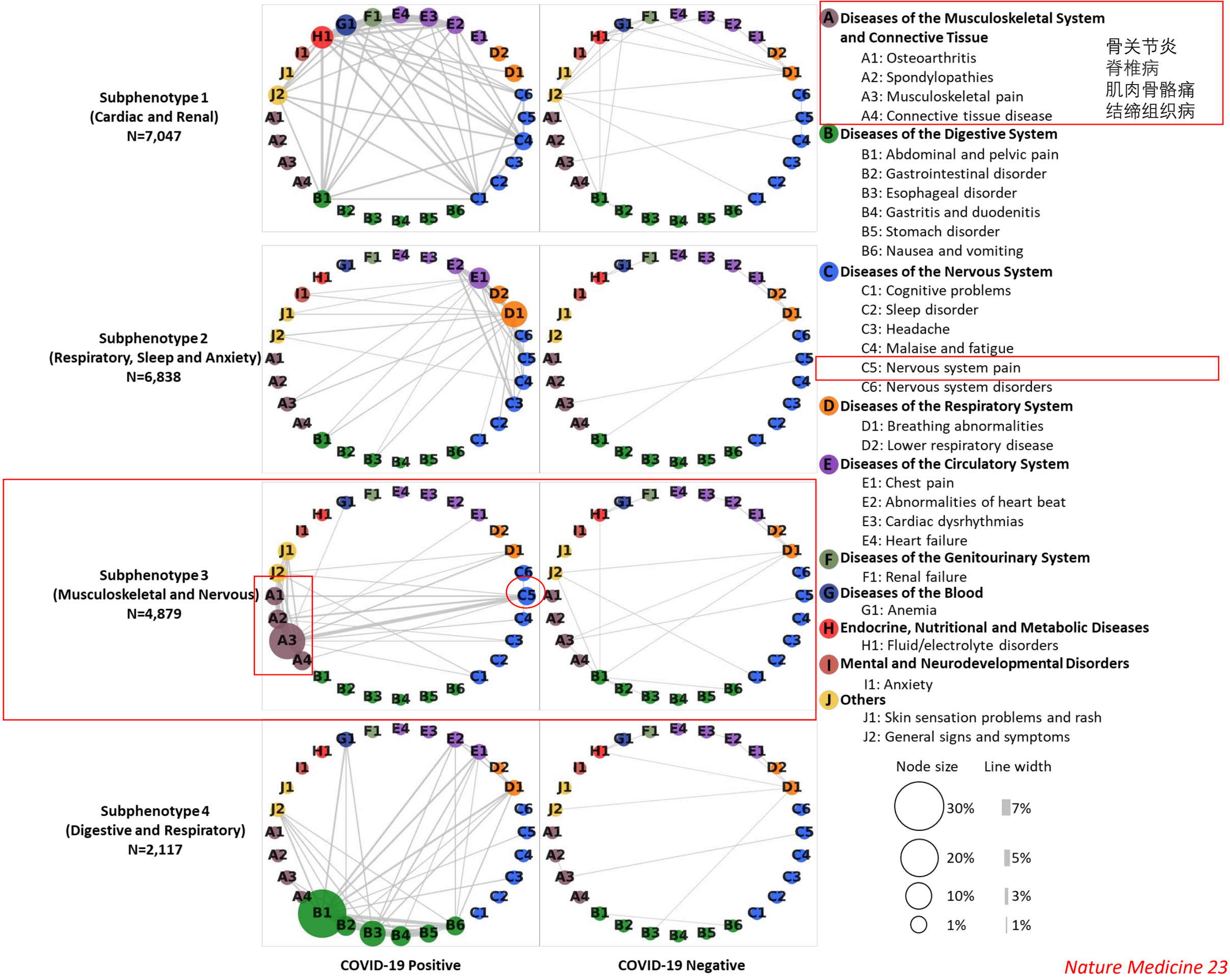
"Data-driven identification of post-acute SARS-CoV-2 infection subphenotypes." *Nature Medicine* 29.1 (2023): 226-235.



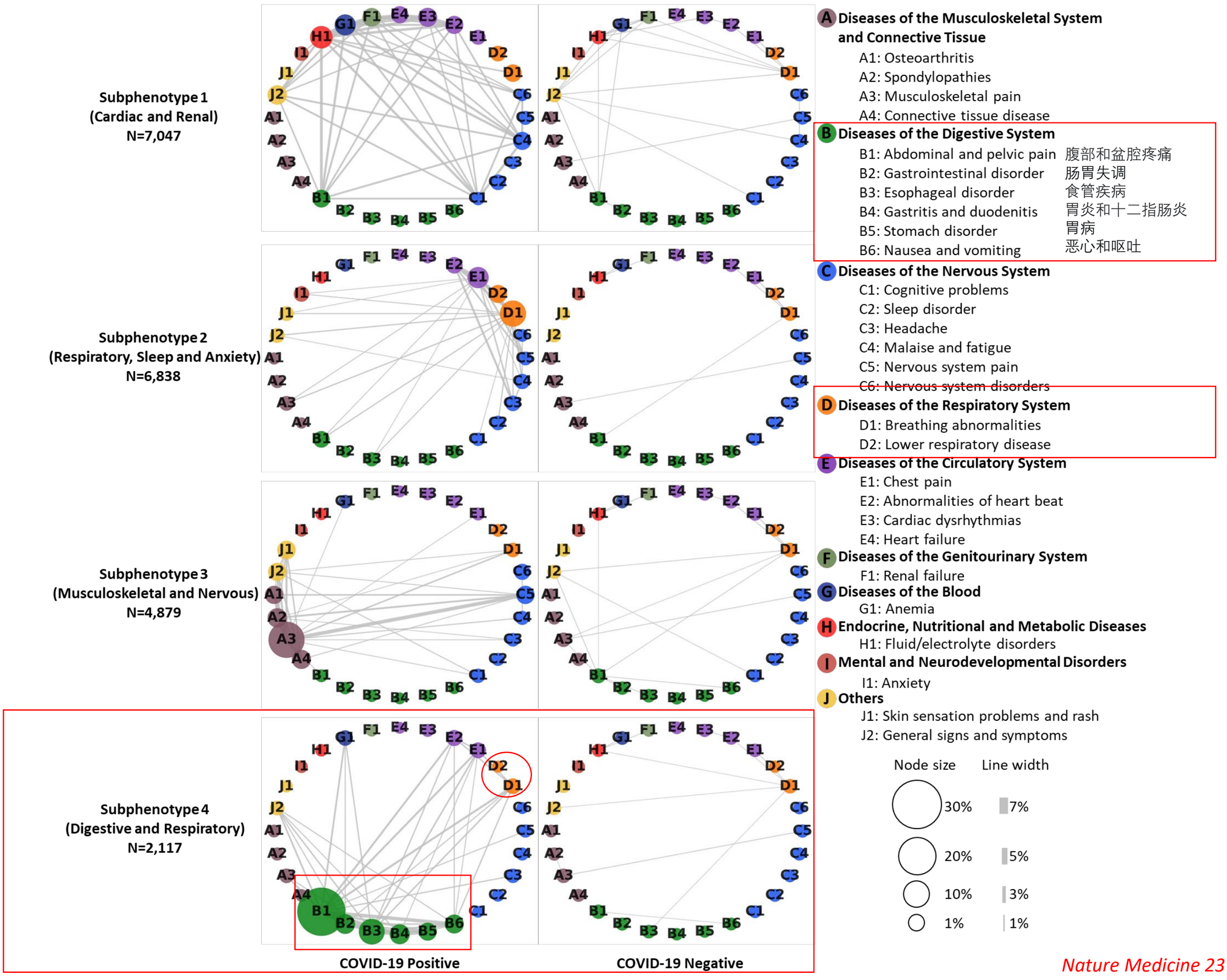
Nature Medicine 23



Nature Medicine 23



Nature Medicine 23



Nature Medicine 23

Clinical implications:

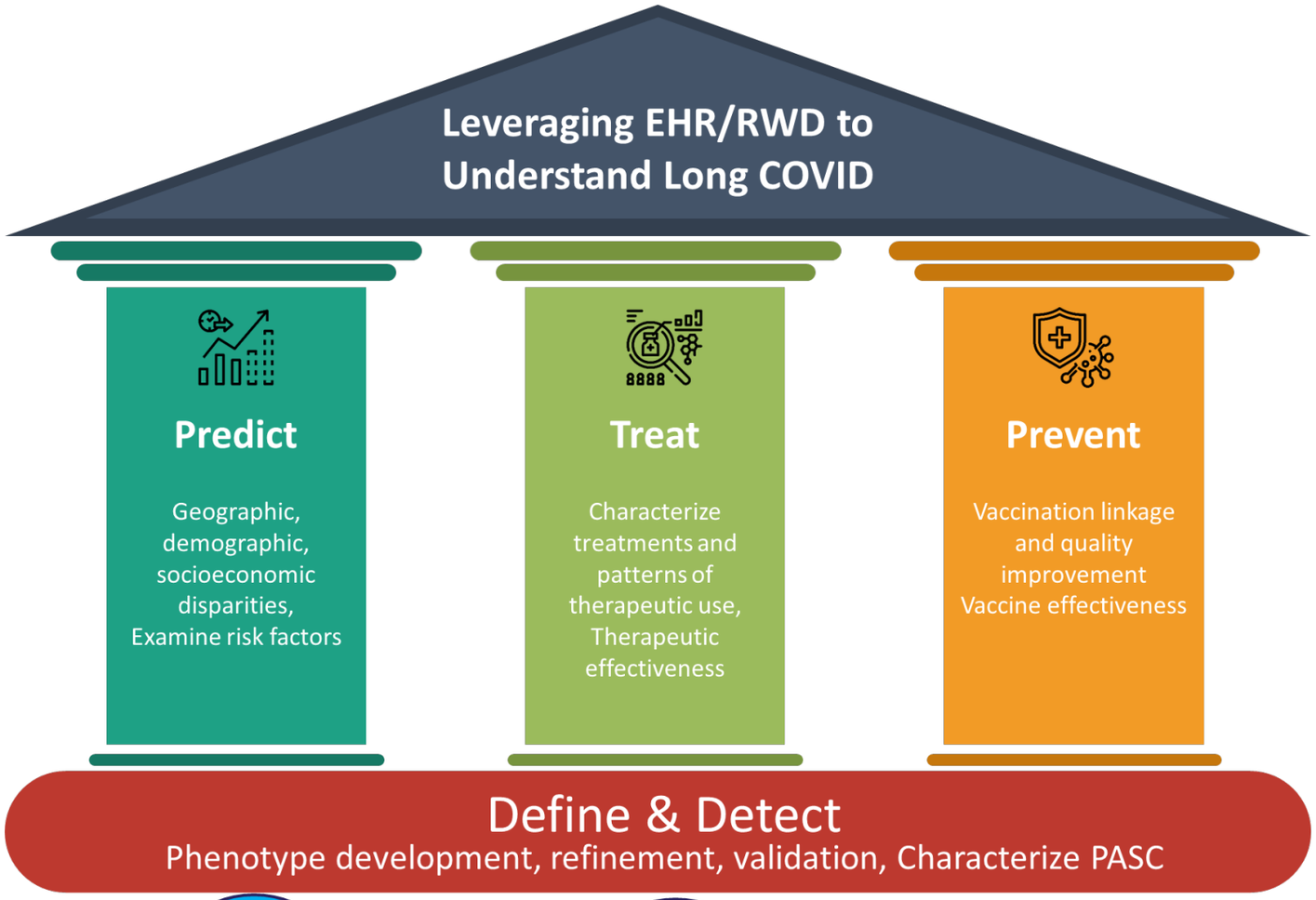
pcornet The National Patient-Centered Clinical Research Network

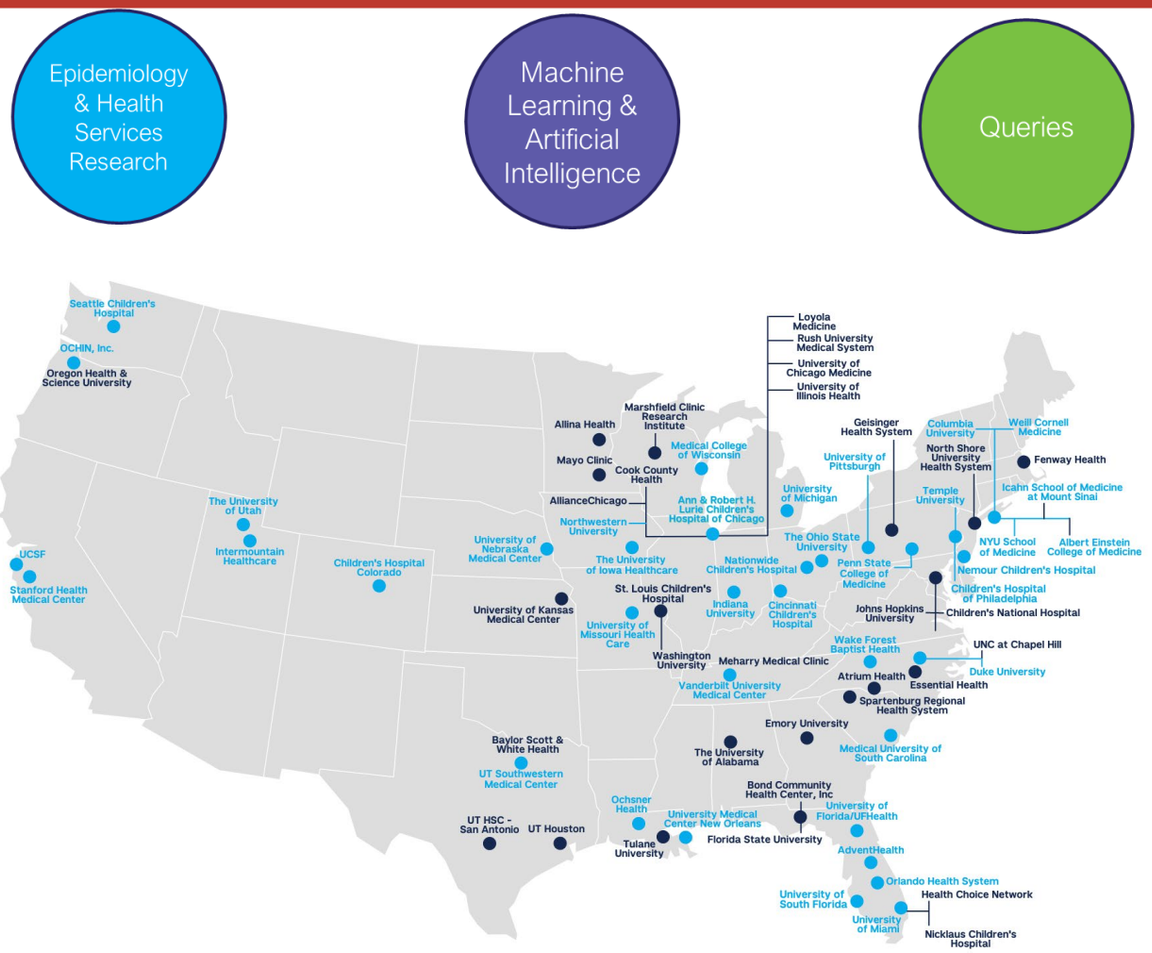
RECOVER Researching COVID to Enhance Recovery An Initiative Funded by the National Institutes of Health

- RWD-driven defining diseases: new & complex
- PASC (or Long COVID) is clinically diverse
  - Incident conditions across different organ systems
  - Geographic variation, may stem from temporal advancements in treatment or variants or different social demographics, etc.
- 4 identified sub-phenotypes
  - Cardiac and renal (median age 65, 51% female, higher severity in acute phase)
  - Respiratory conditions, sleep disorders & anxiety (median age 51, 63% female, lowest rates of hospitalization)
  - Musculoskeletal and nervous system (median age 57, 61% female)
  - Digestive system and respiratory conditions (median age 54, 62% female, lowest rates of ICU care)

1st Takeaway

- RWD/RWE might rapidly improve our understanding of and ability to predict, treat, and prevent Long COVID.
- A natural history to understand and fight against a new emerging disease.

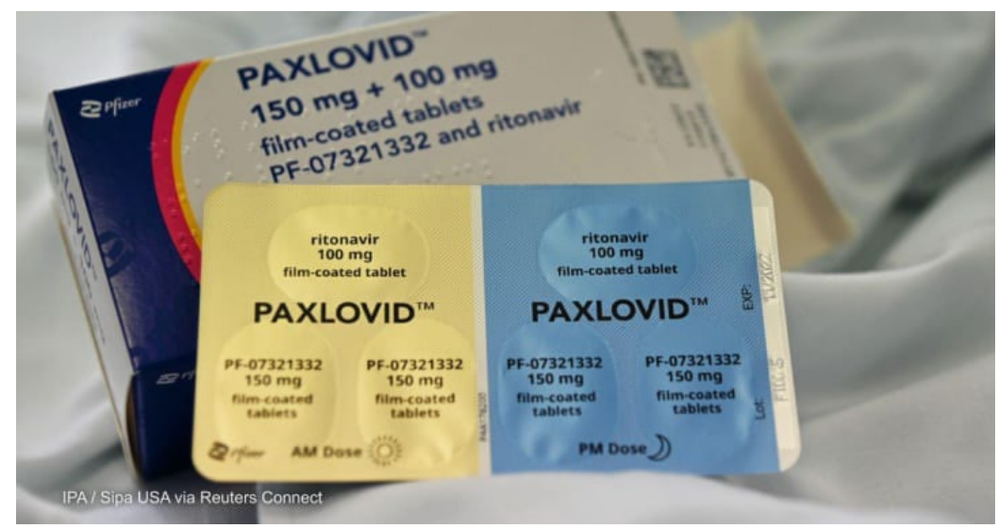




Our ongoing efforts

Predict

Treatment



[Environmental Advances,](#)

[Research Square](#)

2nd Takeaway

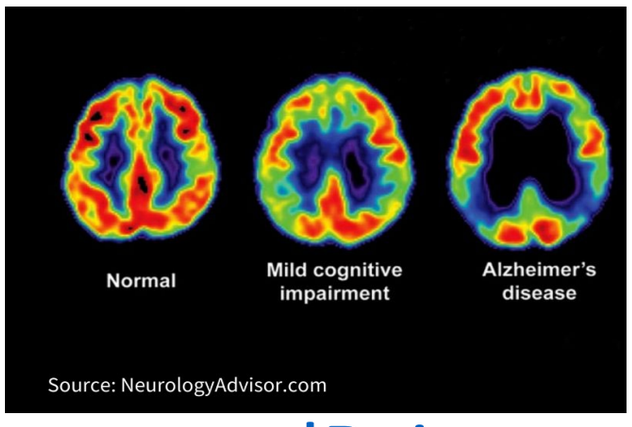
$(X, T, Ys?)$

Increasing throughputs in generating RWE

opens a new door to understanding complex diseases

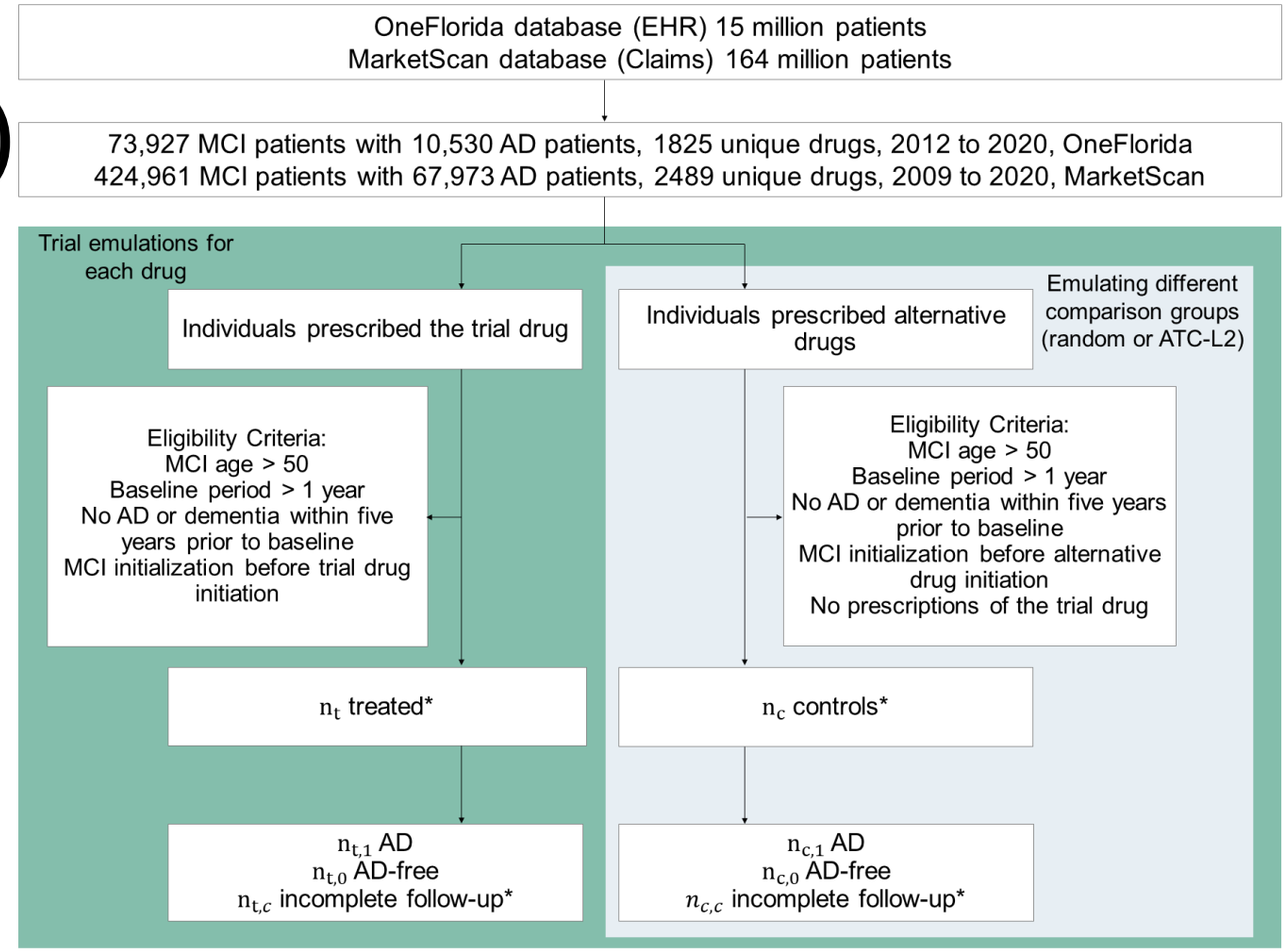
$$(X, T, Ys) \rightarrow (X, Ts?, Y)$$

a new door to RWD-driven drug repurposing

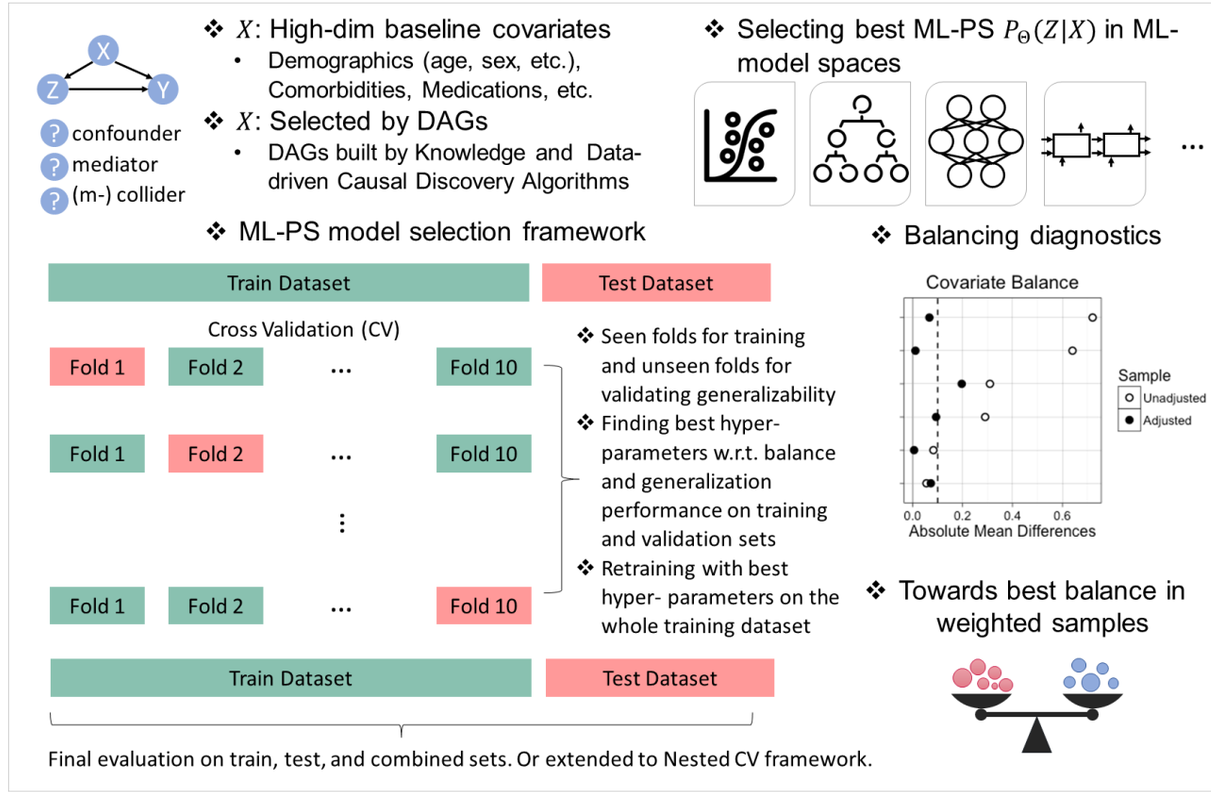


medRxiv

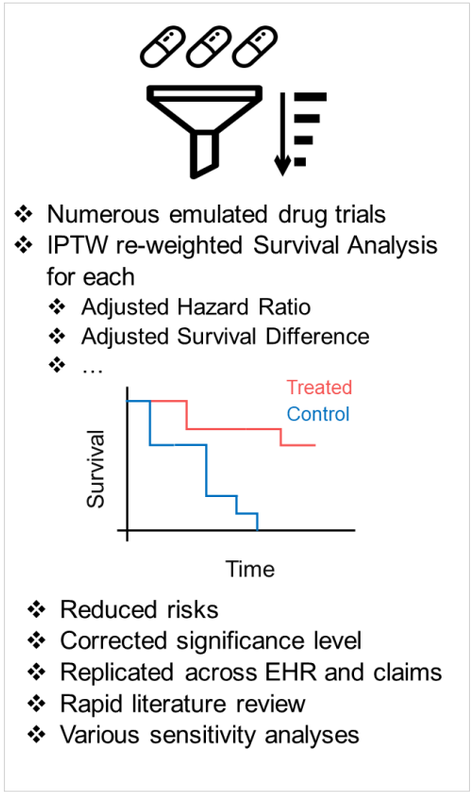
(a) Trial emulations on two RWDs (EHR and Claims)



(b) ML-PS model selection for better balancing performance

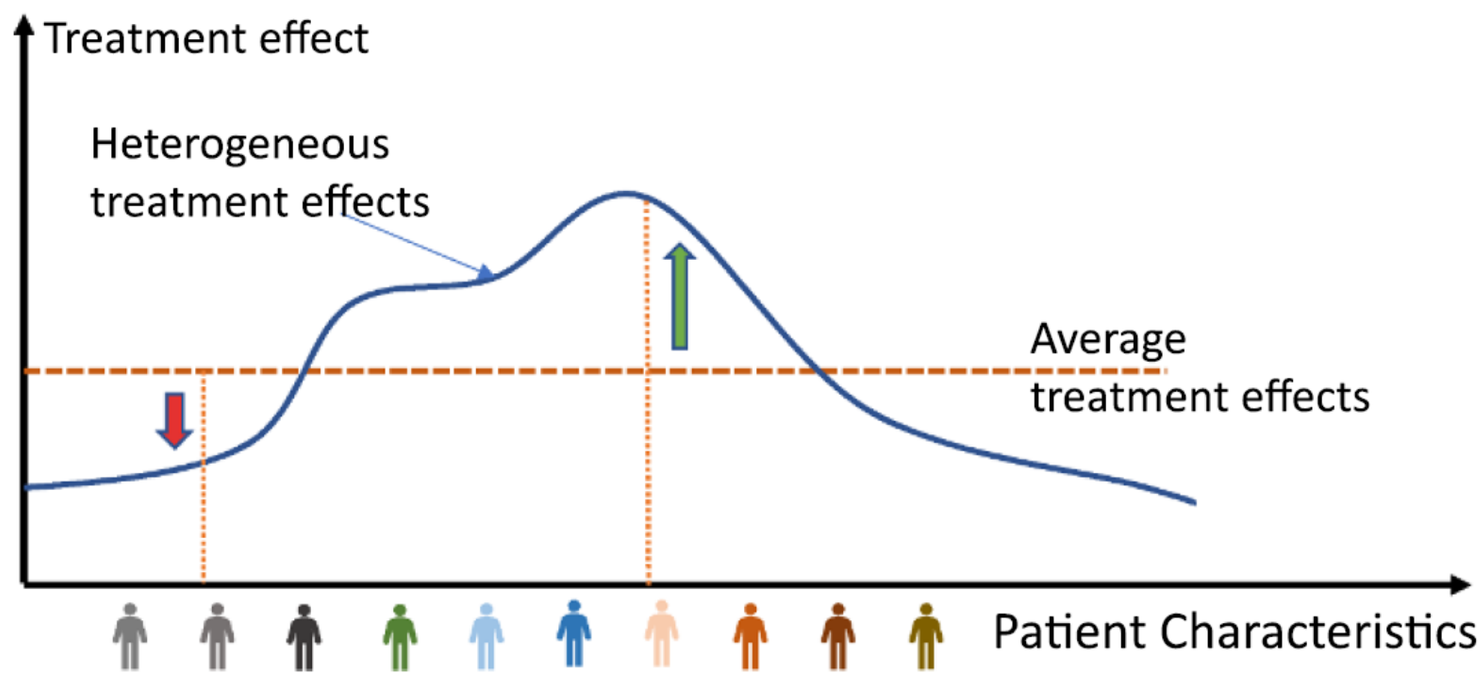


(c) High-throughput screening repurposing drugs



$(X, T, Ys) \rightarrow (X, Ts, Y) \rightarrow (Xs?, T, Y)$

And, a new door to RWD-driven **Trial designs** or Personalized/subtyping treatment



Complex diseases:
ICU - Septic Shock
Chronic – AD
Long COVID

3rd Takeaway

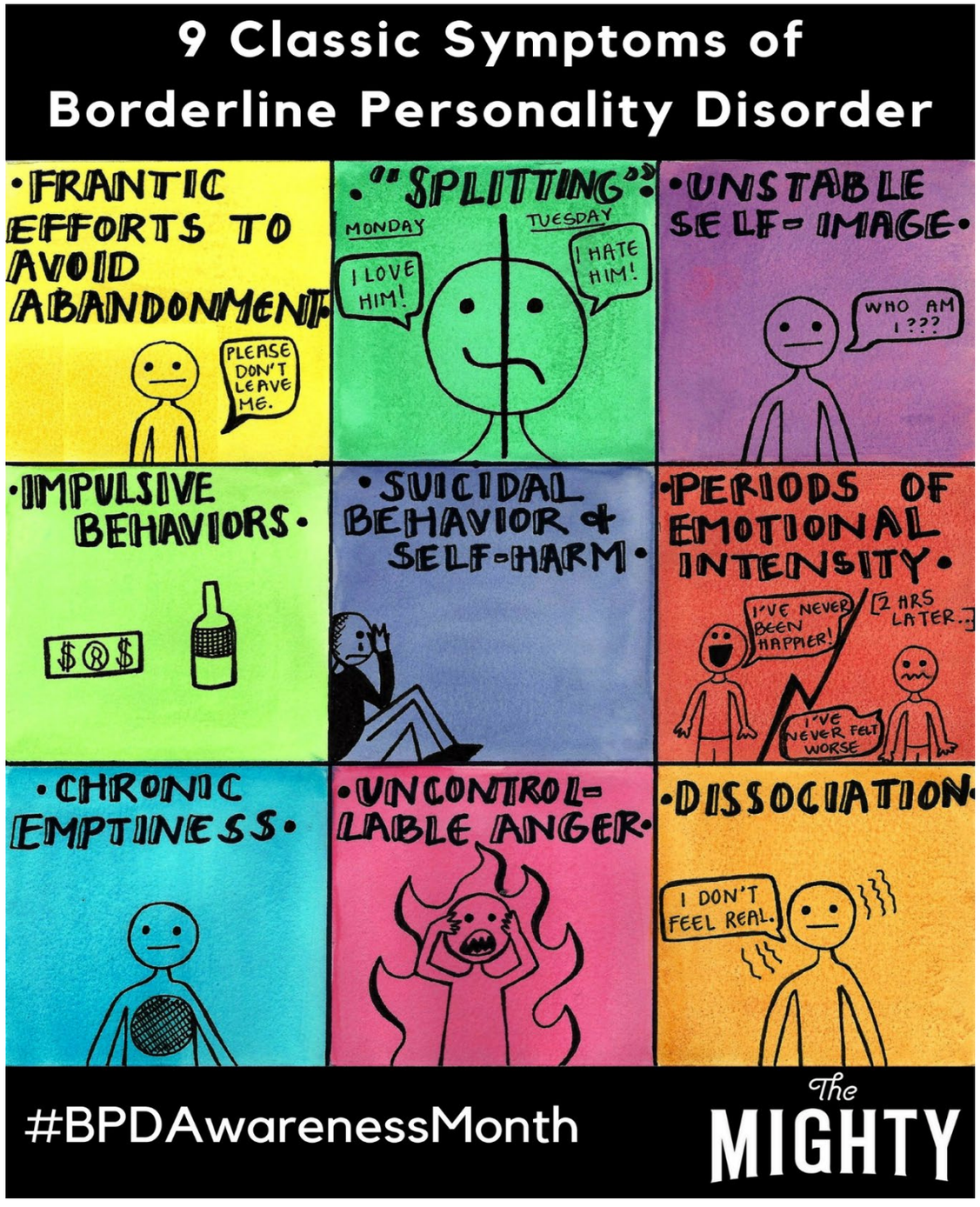
A lot of details and considerations behind the scenes to make the generated evidence more robust and generalizable

- Which covariates should be included?
  - Causal diagrams: Directed acyclic graph (DAG),
  - Clinical knowledge vs. data-driven
- Beyond confounding bias:
  - Confounding by indication
  - Residual confounding
  - Prevalent-user bias
  - Immortal time bias
  - Missing data
  - Misclassification
  - Informative censoring
  - Time-varying exposure
  - Time-varying confounding
- Design:
  - Trial designs
  - Using the right comparators
    - Active comparator?
  - Different diseases
- PS-based methods
  - Assumptions
  - Why it works?
  - Matching
  - Re-weighting
  - …
  - Experiment design vs. outcome models
- Outcome models
  - Meta learners
  - Representation learning
  - Causal forest
  - Doubly robust estimator
- Time-varying exposures
- Evaluation:
  - Balance diagnostics
  - Model selection
  - Cross-validation
- Multiple testing correction
- Simulation
  - Statistical → complex high-dimensional
- Sensitivity analysis
  - Negative control
  - …
- Generalizability
  - Multiple sites
  - Multiple data
- More applications:
  - Drug repurposing?

4th Takeaway

More and More New Usage of RWD/RWE!

RWD-based Screening for Clinical Trial Recruitment



- **Goal:**
  - ***Speed up clinical trials by RWD + AI***
  - **High-throughput Screening** Borderline Personality Disorder patients for Clinical Trial Recruitment (1402-0012) in Boehringer Ingelheim Pharmaceuticals, Inc.
- **Borderline Personality Disorder**
  - A mental illness marked by an ongoing pattern of varying moods, unstable self-image, and behavior, suicidal behavior & self-harm, etc.
- **Challenges: largely Under- or Mis-diagnosed (w/o ICD-10 F60.3)**
  - Not covered by insurance; High rate of comorbid conditions; Negative stigma; Caring cost; No cures;

Boehringer Ingelheim

SciRep 2022

Generate RWE using RWD in different study designs

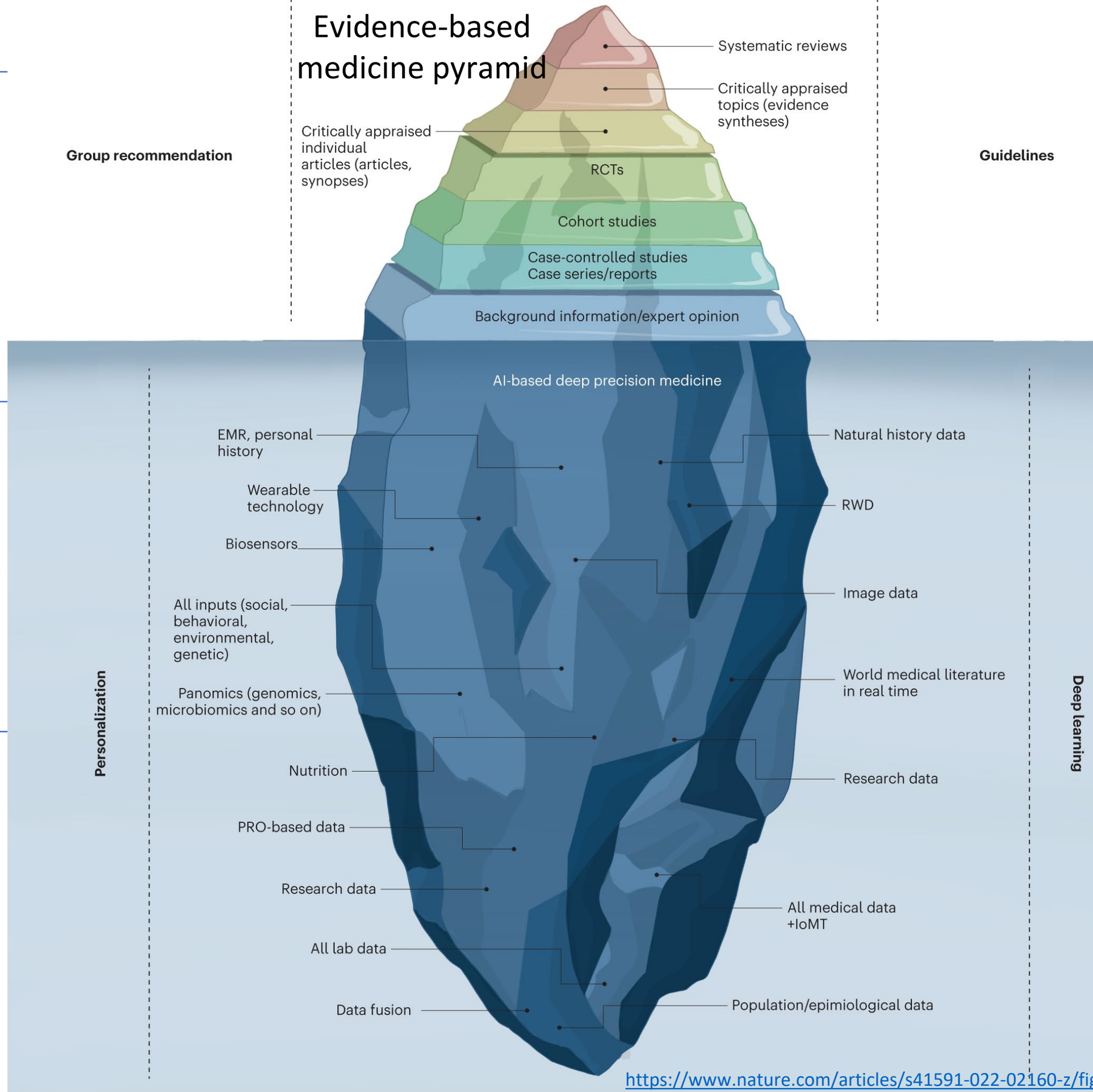
	Non-Randomized	Randomized
Non-Interventional	***Observational Study*** *Cohort study, case-control study, case-crossover study, etc.* *Define disease, incidence/prevalence, surveillance, risk factors, burden, etc.*	
Interventional	***Externally controlled trial*** *Single-group trial with external control group derived from RWD* ***Trial emulation*** *Drug RWE (e.g., effectiveness in the long term, general population, e.g., covid vaccine), drug repurposing, comparative effectiveness, Post-market safety, effectiveness monitoring*	***RCTs using RWD*** *RWD is used to assess enrollment criteria, trial feasibility, recruitment, selection of sites, outcome identification, conduct RCT*

Adapted from NEJM 386.18 (2022): 1680-1682.

FINAL CONCLUSIONS

RWD OFFERS EXCITING OPPORTUNITIES IN GENERATING POTENTIALLY MORE TIMELY/LESS EXPENSIVE/ETHICAL/GENERALIZABLE /COMPREHENSIVE/INDIVIDUALIZED/ HIGH-THROUGHPUT EVIDENCE FOR A WIDE RANGE OF APPLICATIONS

DEEP SYNTHESIS AND INTEGRATION OF THESE DATA NEED INNOVATIONS IN METHODS, APPLICATIONS, SYSTEMS IMPLEMENTATION, AND INTERDISCIPLINARY MINDSETS



https://www.nature.com/articles/s41591-022-02160-z/fig

Selected Media Coverage

- May 2023. Our Data-driven analysis to understand long COVID using electronic health records from the RECOVER initiative was highlighted in Cornell Chronicle: Long COVID risk and symptoms vary across populations and in Weill Cornell Medicine Newsroom: Study Discovers Long COVID Risk and Symptoms Vary in Different Populations
- May 2023. Our previous molecular generative AI paper MoFlow: An Invertible Flow Model for Generating Molecular Graphs was highlighted in Weill Cornell Medicine Population Health Sciences News
- March 21st, 2023. Our Molecule Generative AI model - MoFlow was highlighted by NVIDIA CEO Jensen Huang @ NVIDIA GTC 2023 Keynote and being integrated into NVIDIA BioNeMo Service for AI-driven Drug Discovery! See the exciting moment and inspiring introduction at 48:00 mins at Youtube:GTC 2023 Keynote with NVIDIA CEO Jensen Huang. Also refer to the NVIDIA Developer Technical Blog: Build Generative AI Pipelines for Drug Discovery with NVIDIA BioNeMo Service for more details.
- March 2023. Our Risk Factors and Predictive Modeling for Long Covid paper was highlighted in News Medical: What are the risk factors associated with post-acute SARS-CoV-2 infection?
- March 2023. Our Racial/Ethnic Disparities in Long Covid paper was highlighted in BMJ News: Covid-19: US studies show racial and ethnic disparities in long covid.
- March 2023. Our Environmental risk factors for Long Covid paper was highlighted in NIH Director's Blog: RECOVER: What Clinical Research Comes Next for Helping People with Long COVID.
- February 2023. Our Racial/Ethnic Disparities in Long Covid paper was highlighted in NIH News Releases. NIH RECOVER research identifies potential long COVID disparities.
- February 2023. Our Racial/Ethnic Disparities in Long Covid paper was highlighted in Cornell Chronicle and Weill Cornell Medicine Newsroom: Long COVID Symptoms Vary Among Racial and Ethnic Groups; Cancer Health: RECOVER Research Identifies Potential Long COVID Disparities; and Bet : Black, Hispanic Patients More Likely To Develop Lasting Symptoms After COVID.
- February 2023. Our Long Covid subphenotyping paper was highlighted in NIH - News and Stories: Researchers Identify Four Long COVID Categories.
- Janunary 2023. Our Long Covid subphenotyping paper was highlighted in Cornell Chronicle: Study identifies four major subtypes of long COVID; CN-HEALTHCARE 健康界: Nat Med：研究近3.5万名新冠患者数据，确定了长新冠存在四种主要的症状模式; Medical Xpress: Study identifies four major subtypes of long COVID; Verywellhealth: Long COVID May Manifest Itself in 4 Major Ways, Research Shows; Prevention: Study Finds There Are 4 Subtypes of Long COVID, New Atlas:Four distinct subtypes of long COVID defined in machine learning study; Miami Herald: There are 4 'major' types of long COVID symptoms, study finds. How likely is each?; and BOSTON.com:New study categorizes long COVID symptoms, allowing for earlier detection, "They don't have to suffer in silence."
- Janunary 2023. Our Long Covid subphenotyping paper was highlighted in Nature Medicine - Research Briefing: Machine learning identifies long COVID patterns from electronic health records.
- December 2022. Our Long Covid subphenotyping paper was highlighted in Weill Cornell Medicine Newsroom. Study Identifies Four Major Subtypes of Long COVID, and MedPage Today. Are Subphenotypes for Long COVID Beneficial? — A new study can help physicians evaluate potential treatment approaches.
- June 2022. Our Long Covid subphenotyping paper was highlighted in News Medical. Machine learning analysis suggests that there are four sub-phenotypes of long COVID
- June 2022. Our Long Covid subphenotyping paper was highlighted in Fortune. Long COVID symptoms: What we know—and don't know—about the mysterious illness that could affect up to 80% of COVID survivors
- May 2022. Our Long Covid analysis paper was highlighted in News Medical.Largest study to date on long COVID identifies a broad list of diagnoses.

www.calvinzang.com/#news

References

- **Chengxi Zang**, Yongkang Zhang, Jie Xu, Jiang Bian, Dmitry Morozyuk et al. "Data-Driven Analysis to Understand Long COVID Using Electronic Health Records from the RECOVER Initiative." ***Nature Communication***, 2023.
- Zhang, Hao, **Chengxi Zang**, Zhenxing Xu, Yongkang Zhang, Jie Xu, Jiang Bian, Dmitry Morozyuk et al. "Data-driven identification of post-acute SARS-CoV-2 infection subphenotypes." ***Nature Medicine*** 29, no. 1 (2023): 226-235.
- Zhang, Yongkang, Hui Hu, Vasilios Fokaidis, Jie Xu, **Chengxi Zang**, Zhenxing Xu, Fei Wang et al. "Identifying environmental risk factors for post-acute sequelae of SARS-CoV-2 infection: An EHR-based cohort study from the recover program." ***Environmental Advances*** 11 (2023): 100352.
- Khullar, Dhruv, Yongkang Zhang, **Chengxi Zang**, Zhenxing Xu, Fei Wang, Mark G. Weiner, Thomas W. Carton, Russell L. Rothman, Jason P. Block, and Rainu Kaushal. "Racial/Ethnic Disparities in Post-acute Sequelae of SARS-CoV-2 Infection in New York: an EHR-Based Cohort Study from the RECOVER Program." ***Journal of General Internal Medicine*** (2023): 1-10.
- **Zang, Chengxi**, Hao Zhang, Jie Xu, Hansi Zhang, Sajjad Fouladvand, Shreyas Havaldar, Feixiong Cheng et al. "High-Throughput Clinical Trial Emulation with Real World Data and Machine Learning: A Case Study of Drug Repurposing for Alzheimer's Disease." ***medRxiv*** (2022).
- **Zang, Chengxi**, and Fei Wang. "MoFlow: an invertible flow model for generating molecular graphs." In *Proceedings of the 26th ACM **SIGKDD** International Conference on Knowledge Discovery & Data Mining*, pp. 617-626. 2020.
- Xu, Jie, Fei Wang, **Chengxi Zang**, Hao Zhang, Kellyann Niotis, Ava L. Liberman, Cynthia M. Stonnington et al. "Comparing the effects of four common drug classes on the progression of mild cognitive impairment to dementia using electronic health records." ***Scientific Reports*** 13, no. 1 (2023): 8102.
- **Zang, Chengxi**, Marianne Goodman, Zheng Zhu, Lulu Yang, Ziwei Yin, Zsuzsanna Tamas, Vikas Mohan Sharma, Fei Wang, and Nan Shao. "Development of a screening algorithm for borderline personality disorder using electronic health records." ***Scientific Reports*** 12, no. 1 (2022): 1-12.
- Wanyan, T., Honarvar, H., Jaladanki, S.K., **Zang, C**., Naik, N., Somani, S., De Freitas, J.K., Paranjpe, I., Vaid, A., Zhang, J. and Miotto, R., 2021. Contrastive learning improves critical event prediction in COVID-19 patients. ***Patterns***, *2*(12), p.100389.
- **Zang, Chengxi**, and Fei Wang. "SCEHR: Supervised Contrastive Learning for Clinical Risk Prediction using Electronic Health Records." In *2021 IEEE International Conference on Data Mining (**ICDM**)*, pp. 857-866. IEEE, 2021.

- Nalbandian, Ani, Kartik Sehgal, Aakriti Gupta, Mahesh V. Madhavan, Claire McGroder, Jacob S. Stevens, Joshua R. Cook et al. "Post-acute COVID-19 syndrome." *Nature medicine* 27, no. 4 (2021): 601-615.
- Rosenbaum, Paul R., and Donald B. Rubin. "The central role of the propensity score in observational studies for causal effects." *Biometrika* 70.1 (1983): 41-55.
- Austin, Peter C. "An introduction to propensity score methods for reducing the effects of confounding in observational studies." *Multivariate behavioral research* 46, no. 3 (2011): 399-424.
- Hernán, M. A., & Robins, J. M. (2016). Using big data to emulate a target trial when a randomized trial is not available. American journal of epidemiology, 183(8), 758-764.

Thanks & QA

chz4001@med.cornell.edu
www.calvinzang.com
@calvin_zcx