Integration of Data Streams to Augment Genomic Data

Accelerating the Use of Pathogen Genomics: a workshop

National Academies
Melissa Haendel, PhD, FACMI
University of North Carolina Chapel Hill

On Behalf of the N3C Consortium

These slides: https://bit.ly/nat-academies-n3c

Why N3C?

- Urgent need for observational data at scale
- In the US, there is no centralized healthcare, and therefore no centralized healthcare data
- Data from a single person is spread across multiple providers across time and geography

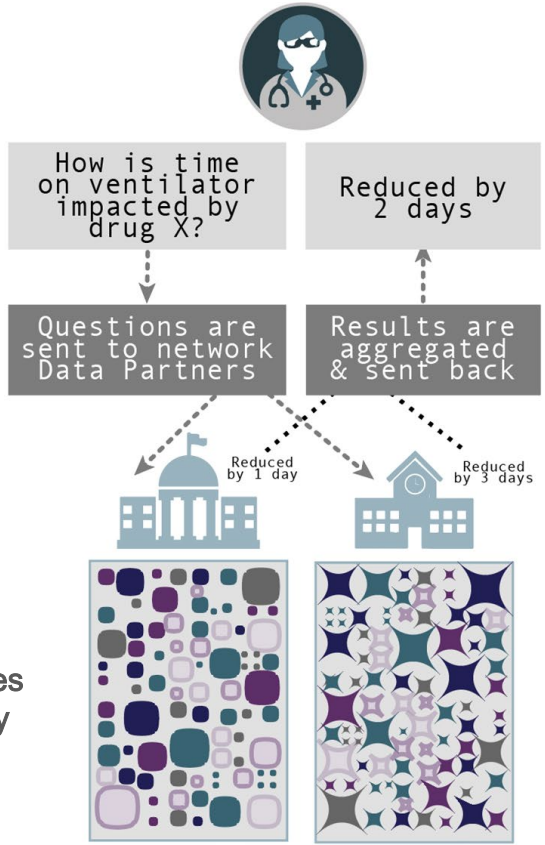
Fragmentation of COVID-19 Patient Clinical Data

National COVID Cohort Collaborative

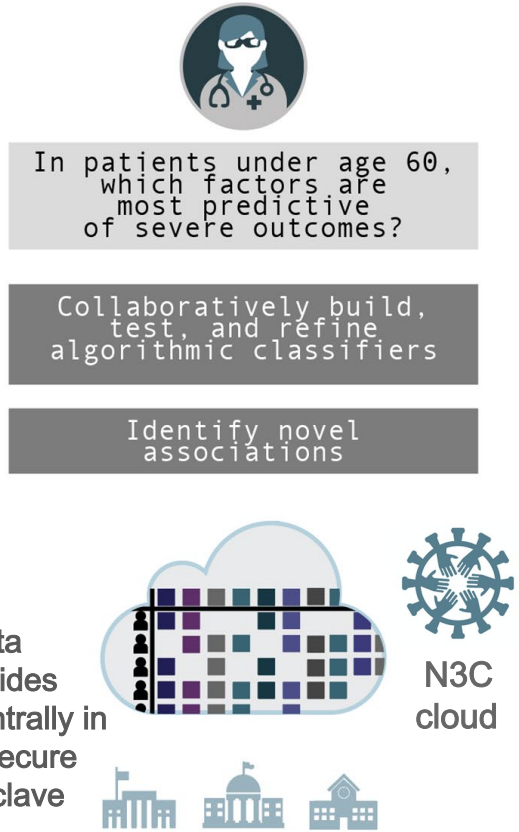
Federated and centralized approaches are synergistic

https://bit.ly/3eMIO1j

Federation is good for specific queries



Centralized analytics are good for discovery



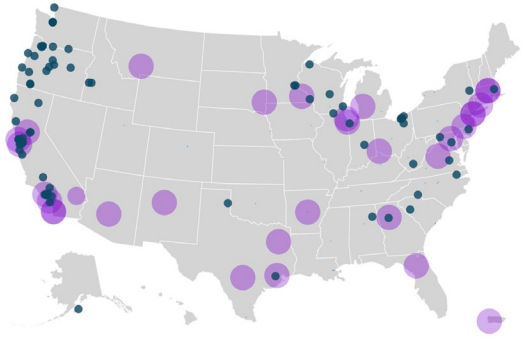
For the new disease that is COVID-19, we needed centralized analytics

National COVID Cohort Collaborative

What is N3C?
National COVID Cohort Collaborative

N3C: largest national public HIPAA-limited data set in US history

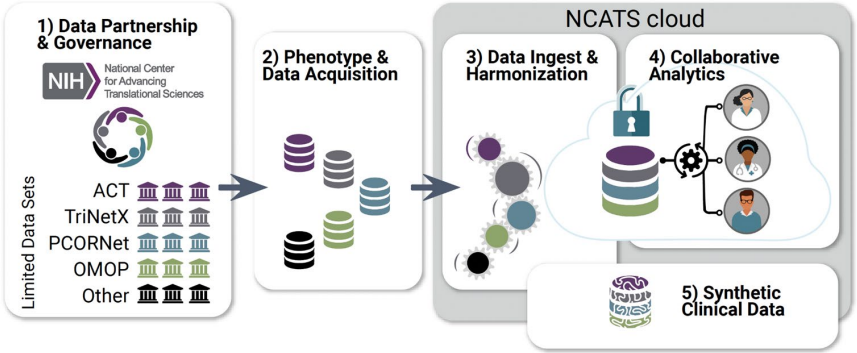
22.8 M records from >230 institutions



covid.cd2h.org/dashboard

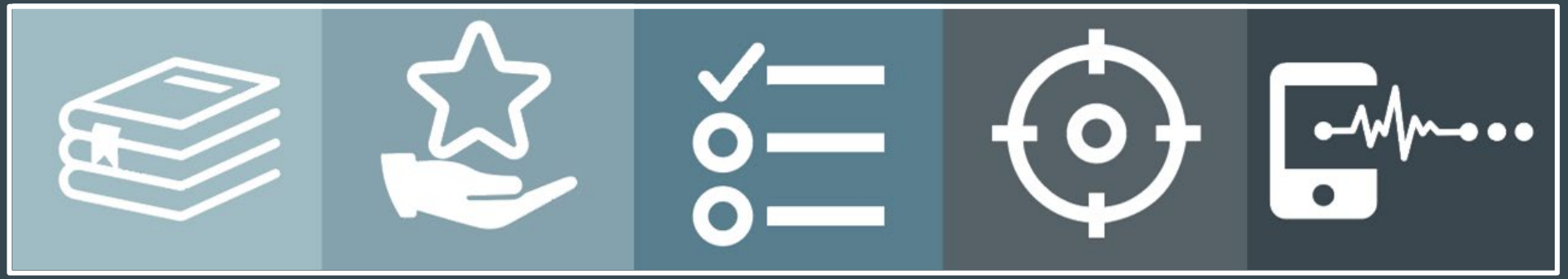
- **Representative pan-US cohort**: race, ethnicity, gender, geography, socio-economic status, health background
- **Harmonized:** overcomes source data heterogeneity
- **Linked:** Patient records, viral variants, vaccine data, CMS, environment, SDoH, etc.
- **Public Health Surveillance:** new variants, comparative effectiveness of drugs

N3C's data-model agnostic, data harmonization and QC pipeline



>450 organizations participating, >4,600 users

Diverse impact of N3C collaborative analytics



RESULTED IN SIGNIFICANT SCHOLARLY PRODUCTIVITY

ATTRIBUTED AT SCALE AND INCENTIVIZED COLLABORATION

TRANSFORMED CARE GUIDELINES

DEVELOPED EVIDENCE-BASED DISEASE DEFINITIONS

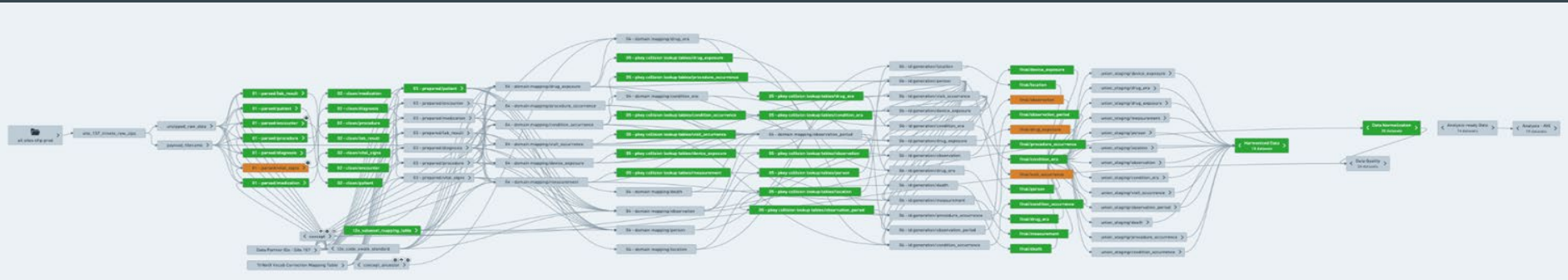
DEVELOPED COMPLEX RISK PREDICTION MODELS

bit.ly/n3c-google-scholar >4300 citations, H index of 30

National COVID Cohort Collaborative

Ensuring harmonization rigor with scalable, continuous monitoring and provenance

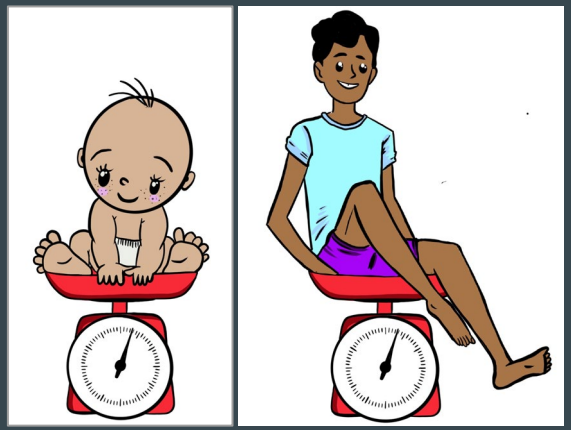
- **Complete transparency into lineage/provenance of harmonization pipelines for >232 sites (77 DTAs) and >50,000 transforms**
- **Pipeline versioning, deployment, upgrades, and automated data quality checks of new and existing sites**
- **Curators and developers can quickly identify and address issues**
- **Scalability of compute resources; pipelines can be refreshed in <20 mins**



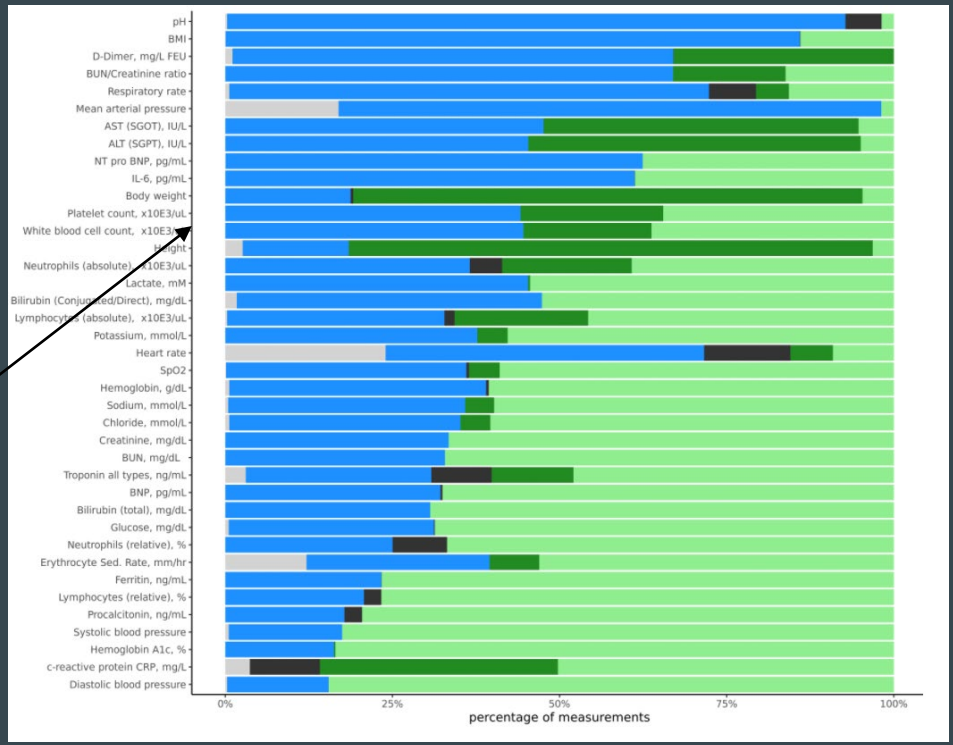
National COVID Cohort Collaborative

Algorithmic data repair is made possible by having many sites

Humans measured in grams do not look the same as humans measured in kilograms!



Canonical unit
Uses a known conversion
Unit not plausible
Missing unit inferred
Unit still missing



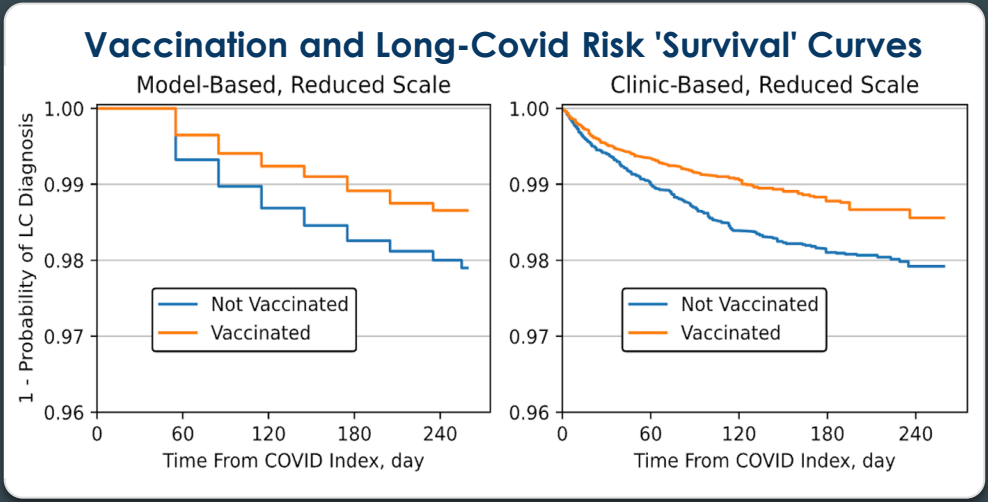
~2x increase in usable data!

https://academic.oup.com/jamia/article/29/7/1172/6569865

N3C provided the earliest and most representative data to predict risk and inform health policy

Problem:

- Conflicting research on effects of vaccination on long-covid
- Determining who is vaccinated is challenging in the US

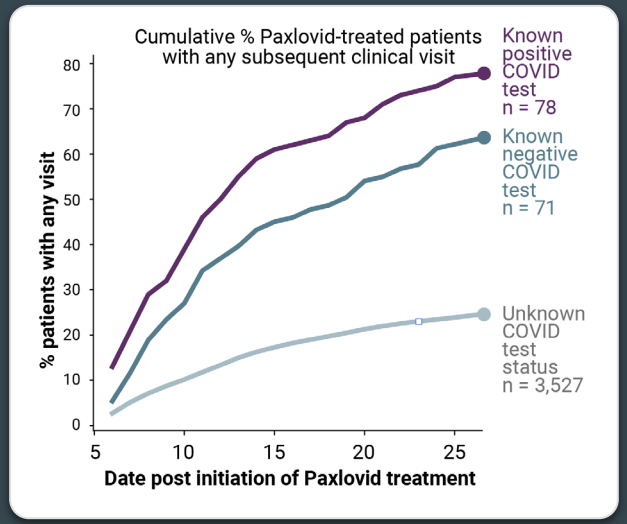


Solution: N3C reconciled vaccination data and demonstrated using multiple methods that vaccination lowers risk of Long-COVID

Nature Comms: https://doi.org/10.1038/s41467-023-38388-7

Problem:

- To plan pandemic response White House needed real-world evidence that Paxlovid was effective



Solution: N3C showed that few patients require care or are hospitalized post-COVID following Paxlovid treatment

https://doi.org/10.1101/2024.01.20.243

National COVID Cohort Collaborative

First evidence-based definition of Long-COVID: a machine learning algorithm identifies PASC patients before they are diagnosed

RECOVER
Researching COVID to Enhance Recovery

An Initiative Funded by the National Institutes of Health

1) EHRs for patients diagnosed with PASC*

**U09.9 code or long-covid clinic*

Machine learning

2) Learned patterns of clinical features of PASC

- Dyspnea
- Fatigue
- No vax on record
- New albuterol Rx
- Many outpatient visits
- New corticosteroid Rx
- …

Search EHRs for similar patients

3) Identify previously unknown cases using learned patterns

The N3C algorithm can be used to identify cohorts for study recruitment and treatment considerations nationwide, in EHRs beyond N3C

Lancet Digital Health; May 16, 2022; https://doi.org/10.1016/S2589-7500(22)00048-6

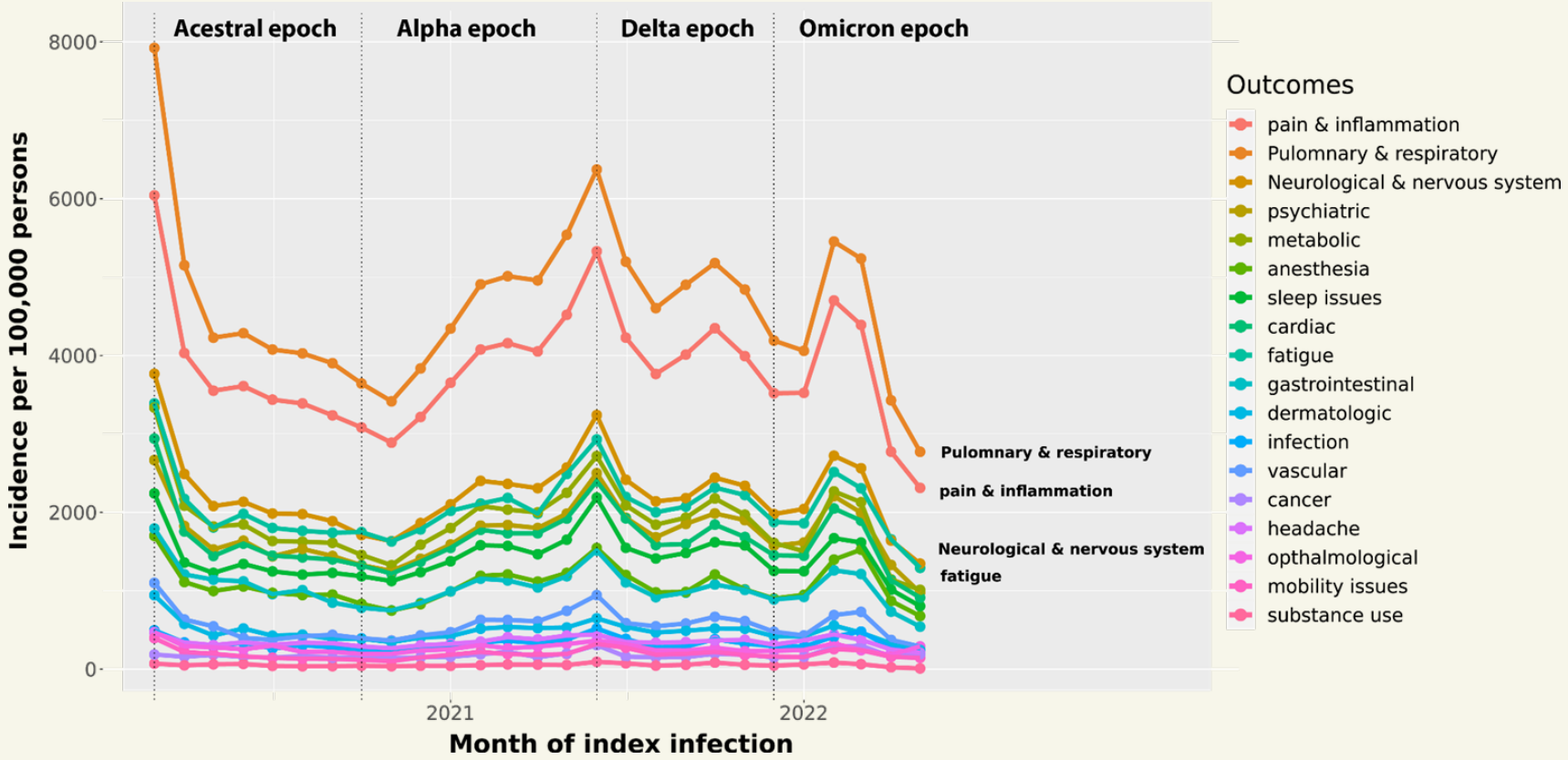
NIH Director's blog: https://directorsblog.nih.gov/tag/national-covid-cohort-collaborative/

National COVID Cohort Collaborative

Important features for classifying patients as having Long-Covid during different viral epochs

RECOVER
Researching COVID to Enhance Recovery

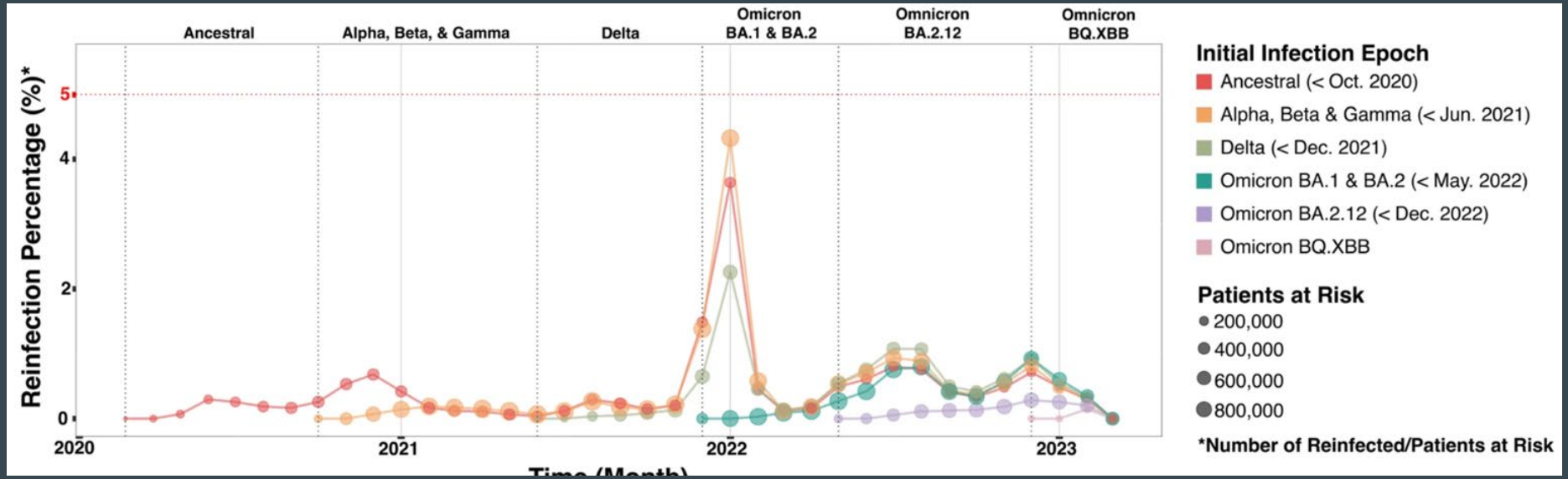
An Initiative Funded by the National Institutes of Health



National COVID Cohort Collaborative

Reinfections during different variant epochs

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



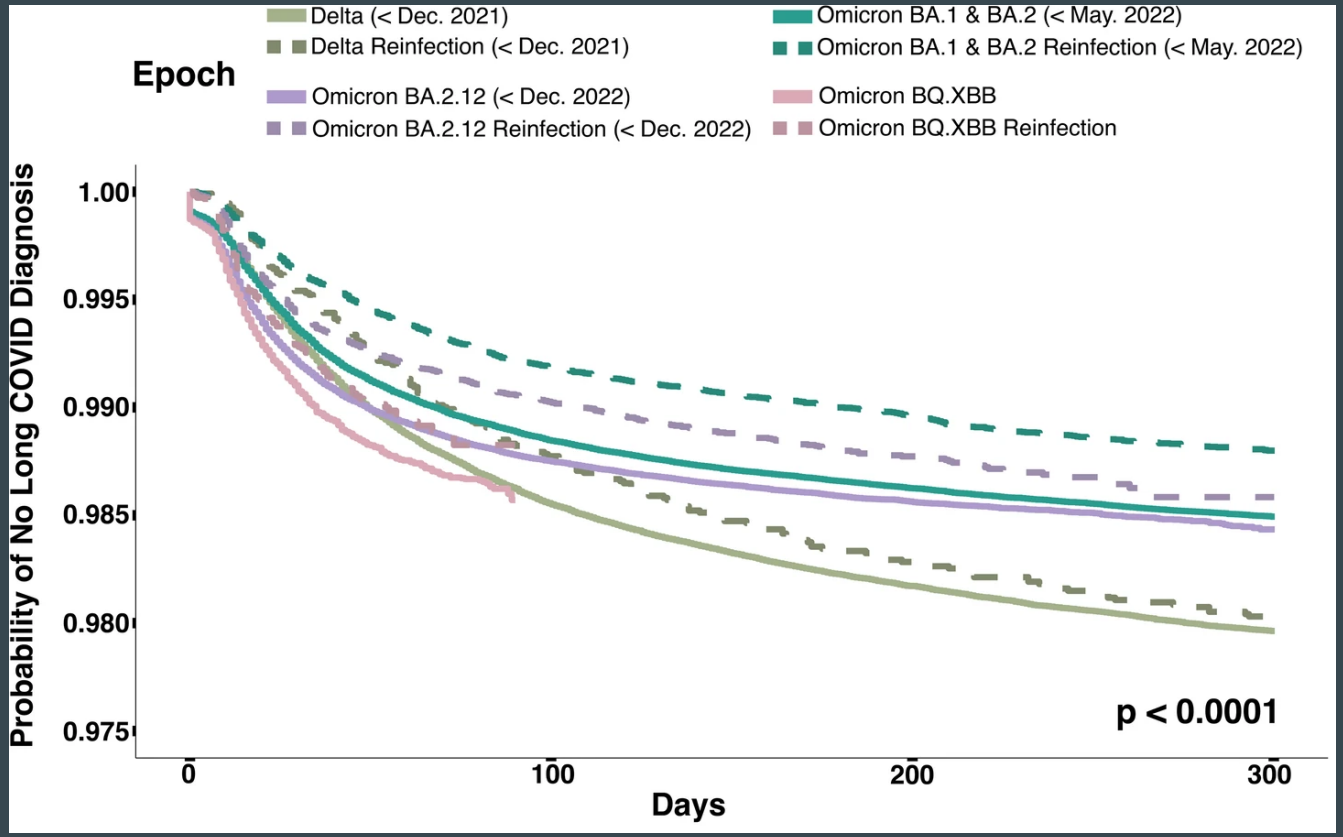
Majority of reinfections occurred during the Omicron epoch

Comms Medicine https://doi.org/10.1038/s43856-024-00539-2

National COVID Cohort Collaborative

Probability of Long-Covid after initial versus reinfection during different variant epochs

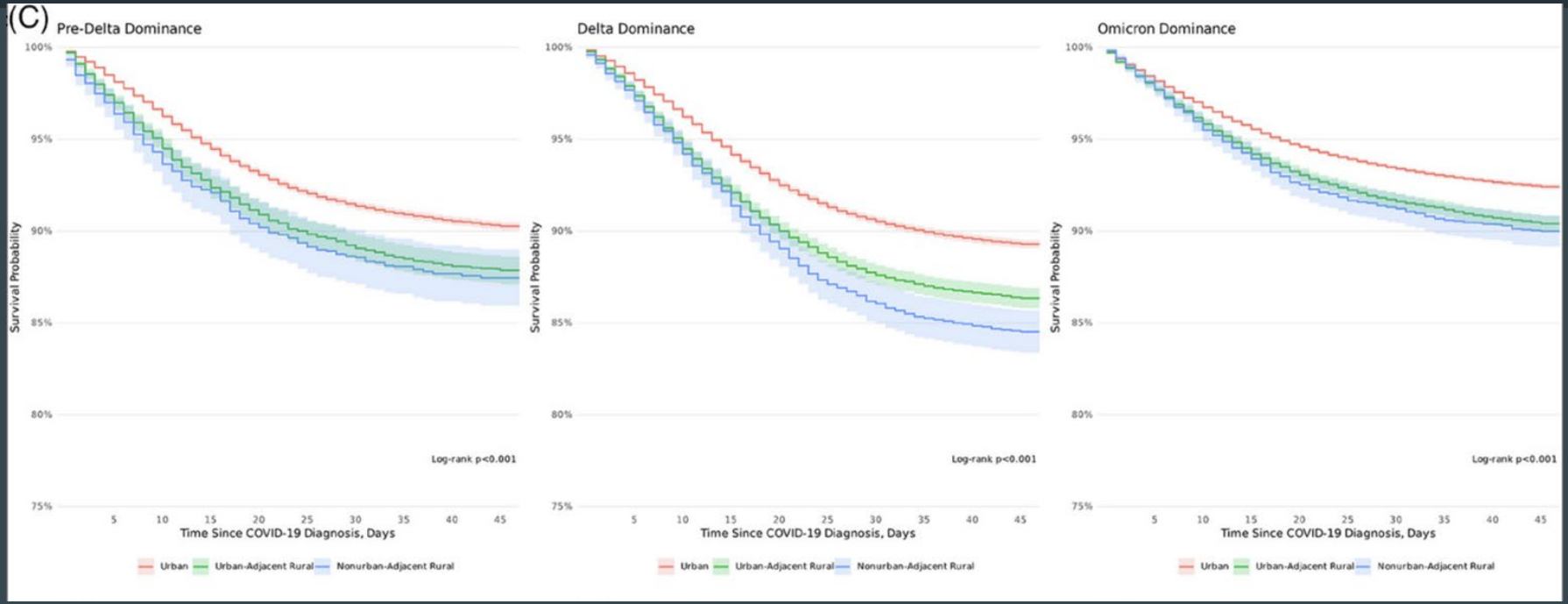
RECOVER
Researching COVID to Enhance Recovery
An Initiative Funded by the National Institutes of Health



- Severity of reinfection associated with severity of initial infection
- Long-COVID diagnoses occur more often following initial infection than reinfection in the same epoch

National COVID Cohort Collaborative

Rural disparities persisted throughout all variant epochs



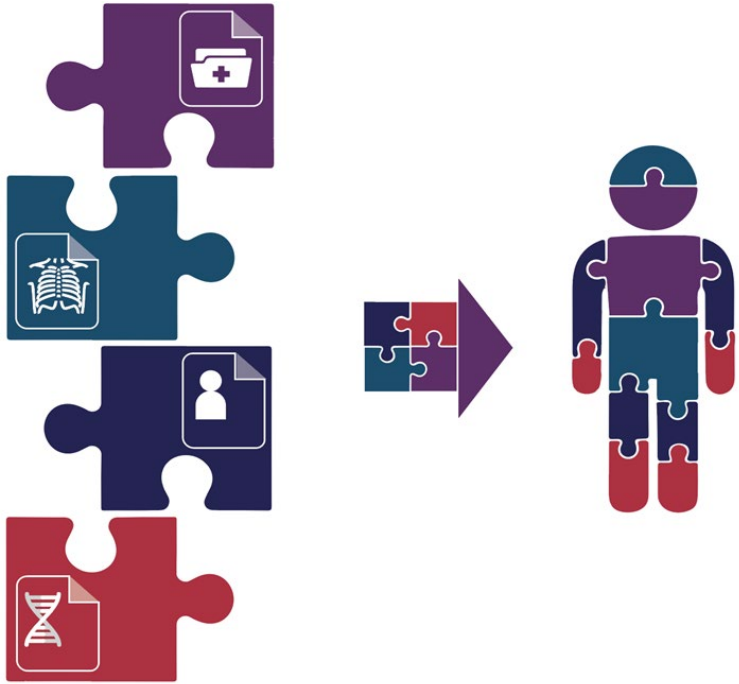
Rural patients had higher hospitalization odds, greater inpatient death hazard, and greater risk of other adverse events compared to urban dwellers. Effectiveness of some therapeutics varied based on rurality.

Journal of Rural Health: https://doi.org/10.1111/jrh.12857

Putting the Patient back together again: EHR data is necessary, but not sufficient

PPRL
- Deduplication
- Multimodal Data
- Cohort Discovery

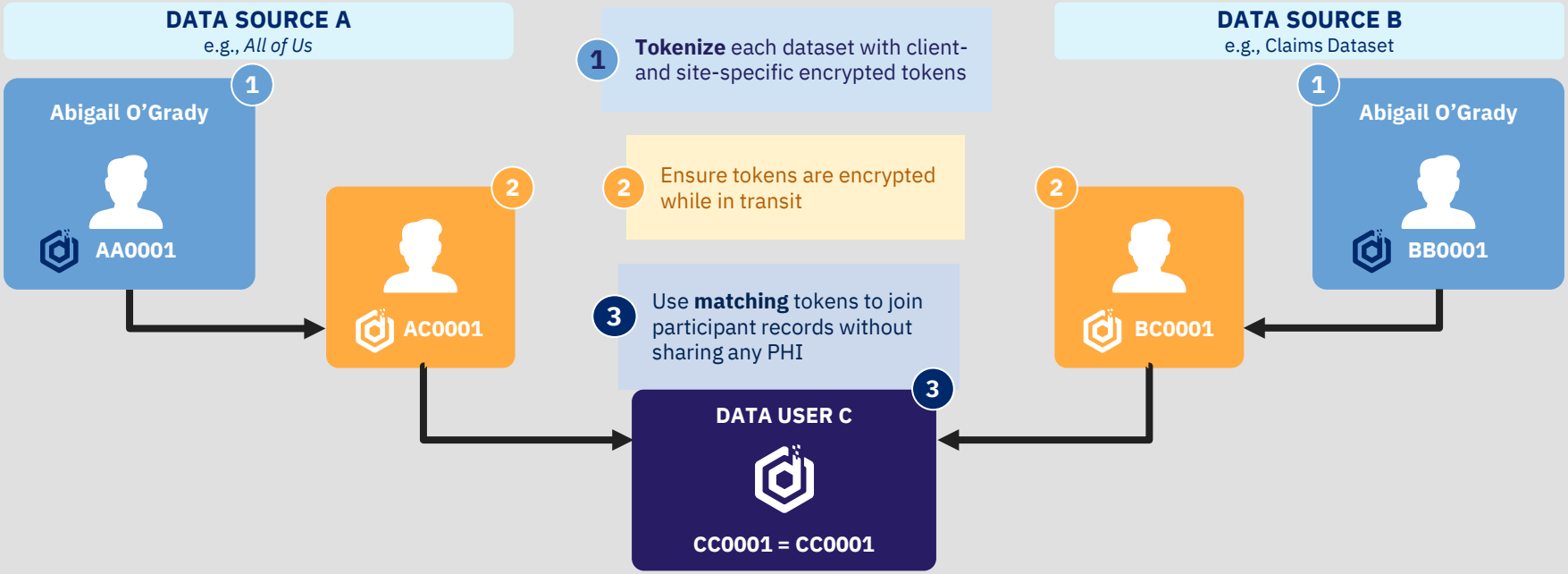
- CMS Data
- SDOH Data
- Viral Variant Mortality Data
- Clinical Data
- Vaccine data
- Imaging
- Registries



Privacy Preserving Record Linkage (PPRL) securely connects records across different data sources while maintaining privacy

National COVID Cohort Collaborative

PPRL meets the definition of de-identified data under HIPAA



What does a token actually look like?

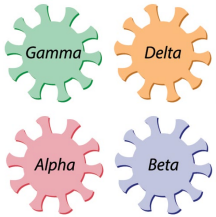
This is just one! We use multiple of these

j05G69K7BH0ugG3VbSusHLbeL7sNm6H3xpiwM6DHfjI=

National COVID Cohort Collaborative

Secure PPRL from EHR data to other data

Non-EHR Mortality Data



Patient Viral Variants

CMS
CENTERS FOR MEDICARE & MEDICAID SERVICES

Medicare/Medicaid Billing Data

SEER

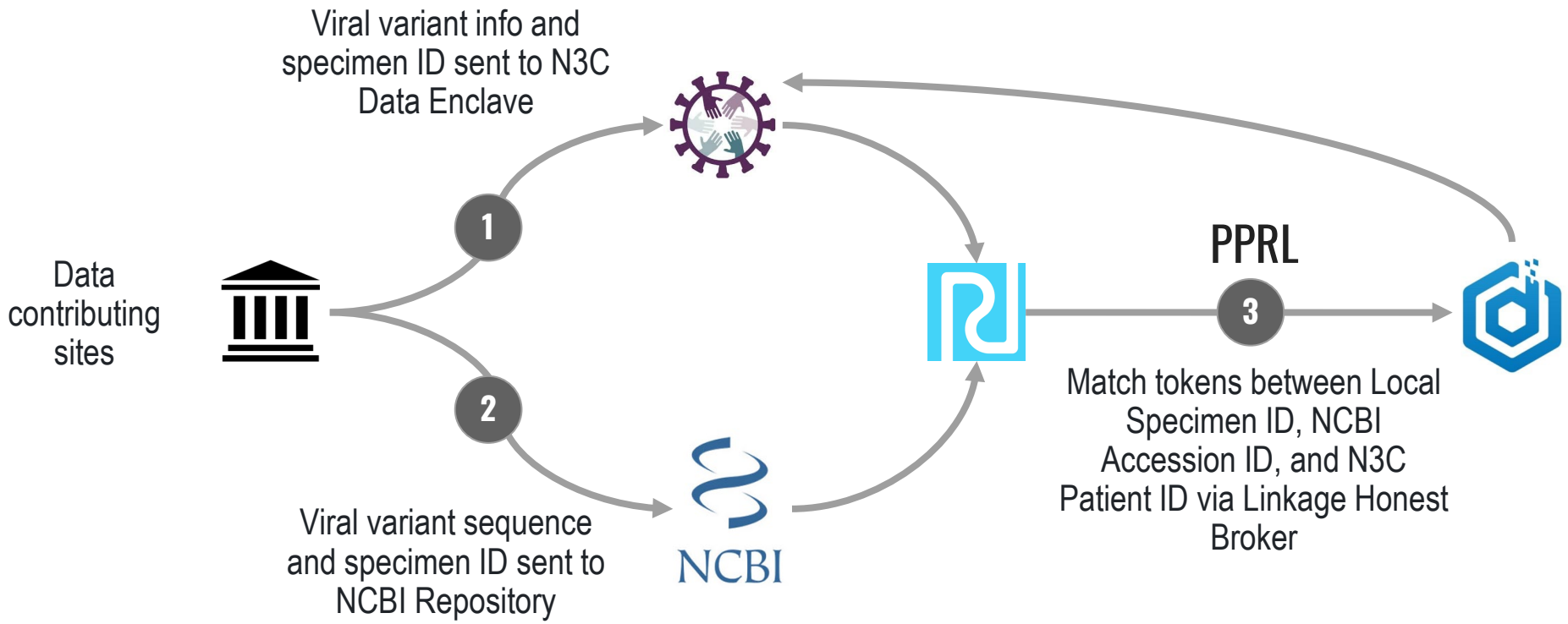
Cancer Registry Data

https://national-covid-cohort-collaborative.github.io/guide-to-n3c-v1/

The Researcher's Guide to N3C

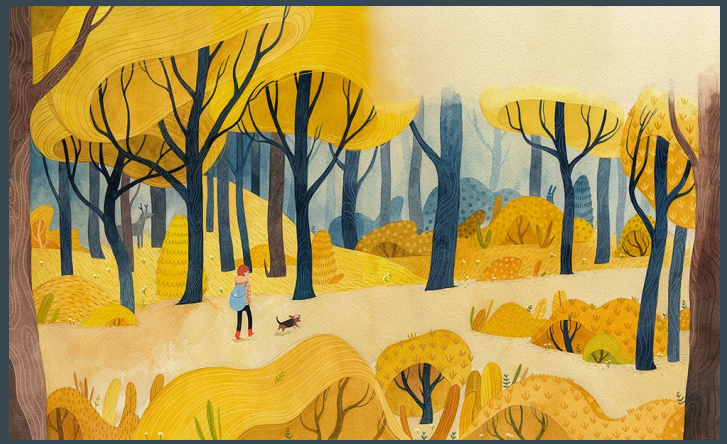
A National Resource for Analyzing Real-World Health Data

Sending viral variants collected from patients to N3C should be easier



github.com/National-COVID-Cohort-Collaborative/variants/wiki

Roads not taken: challenges in linking clinical outcomes



- **EHR data from rural and small clinics very hard to obtain**
  - Lack of representativeness
- **No patient genomic or other ‘omic data to link to**
  - Lack of understanding patient response to infection and modifier variants
- **Vaccination data very hard to obtain and trust**
  - Incomplete vaccination record makes it hard to understand clinical impacts
- **Viral variant data incomplete and challenging to link to patient records**
  - Temporality used more than actual known variants to infer correlations

Image credit: https://www.behance.net/gallery/78642157/Road-Not-Taken

Thank you!

National Center for Advancing Translational Sciences

CTSA Clinical & Translational Science Awards Program

IDeA CTRnet Institutional Development Award — Clinical & Translational Research Network

RECOVER Researching COVID to Enhance Recovery

An Initiative Funded by the National Institutes of Health

The National COVID Cohort Collaborative (N3C): Rationale, Design, Infrastructure, and Deployment

Journal of the American Medical Informatics Association, ocaa196,
https://doi.org/10.1093/jamia/ocaa196
Published: 17 August 2020 **Article history** ▾

bit.ly/n3c-methods-jamia

OPEN ACCESS

Melissa A. Haendel,[1,4,7,8,10,13,14,52,78,101] Christopher G. Chute,[1,4,8,10,13,14,52,78,100,101] Tellen D. Bennett,[9,10,13,14,52,100,101] David A. Eichmann,[4,9,10,13,78,101] Justin Guinney,[4,9,10,14,78,101] Warren A. Kibbe,[9,10,52,78,101] Philip R.O. Payne,[4,9,10,78,101] Emily R. Pfaff,[9,10,13,15,52,78] Peter N. Robinson,[4,9,10,15,52,78,100] Joel H. Saltz,[10,13,14,15,52,78,101] Heidi Spratt,[9,10,100] Christine Suver,[10,78,101] John Wilbanks,[10,78,101] Adam B. Wilcox,[10,101] Andrew E. Williams,[10,13,78] Chunlei Wu,[9,13,14,78] Clair Blacketer,[15,52] Robert L. Bradford,[9,52] James J. Cimino,[10,14,101] Marshall Clark,[9,15,52] Evan W. Colmenares,[9,15,52] Patricia A. Francis,[78] Davera Gabriel,[9,10,13,14,15,52] Alexis Graves,[7,9,78] Raju Hemadri,[9,15,52] Stephanie S. Hong,[9,15,52] George Hripscak,[10,52] Dazhi Jiao,[9,15,52] Jeffrey G. Klann,[14,52,101] Kristin Kostka,[9,15,52] Adam M. Lee,[9,15,52] Harold P. Lehmann,[9,15,52] Lora Lingrey,[9,15,52] Robert T. Miller,[9,15,52] Michele Morris,[9,15,52] Shawn N. Murphy,[9,15,52] Karthik Natarajan,[9,15,52] Matvey B. Palchuk,[9,15,52] Usman Sheikh,[9,78] Harold Solbrig,[9,15,52] Shyam Visweswaran,[10,15,52,101] Anita Walden,[7,10,13,14,52,101] Kellie M. Walters,[10,14,101] Griffin M. Weber,[10,101] Xiaohan Tanner Zhang,[9,15,52] Richard L. Zhu,[9,15,52] Benjamin Amor,[78] Andrew T. Girvin,[15,78] Amin Manna,[78] Nabeel Qureshi,[15,78] Michael G. Kurilla,[10,78] Sam G. Michael,[10,78] Lili M. Portilla,[101] Joni L. Rutter,[1,101] Christopher P. Austin,[101] Ken R. Gersing,[78,101]

Shaymaa Al-Shukri,[4,15] Adil Alaoui,[101] Ahmad Baghal,[15] Pamela D. Banning,[15,100] Edward M. Barbour,[8,15] Michael J. Becich,[15,52,101] Afshin Beheshti,[14] Gordon R. Bernard,[8,15] Sharmodeep Bhattacharyya,[100] Mark M. Bissell,[9,15] L. Ebony Boulware,[14,100] Samuel Bozzette,[100,101] Donald E. Brown,[101] John B. Buse,[14] Brian J. Bush,[8,101] Tiffany J. Callahan,[14,52] Thomas R. Campion,[8,15] Elena Casiraghi,[9,15] Ammar A. Chaudhry,[13,14] Guanhua Chen,[9] Anjun Chen,[13] Gari D. Clifford,[8,15] Megan P. Coffee,[14,100] Tom Conlin,[14] Connor Cook,[7,78] Keith A. Crandall,[9,14,101] Mariam Deacy,[78] Racquel R. Dietz,[78] Nicholas J. Dobbins,[8,9] Peter L. Elkin,[15,52,100] Peter J. Embi,[52,101] Julio C. Facelli,[8,15] Karamarie Fecho,[13] Xue Feng,[9] Randi E. Foraker,[8,13,15] Tamas S. Gal,[8,15] Lingiang Ge,[14] George Golovko,[15,101] Ramkiran Gouripeddi,[14,15] Casey S. Greene,[13,14] Sangeeta Gupta,[52,101] Ashish Gupta,[13,101] Janos G. Hajagos,[9,15] David A. Hanauer,[15,52] Jeremy Richard Harper,[9,14,52] Nomi L. Harris,[14] Paul A. Harris,[101] Mehadi R. Hassan,[9] Yongqun He,[15,52,100] Elaine L. Hill,[9,14] Maureen E. Hoatlin,[14] Kristi L. Holmes,[4,101] LaRon Hughes,[14] Randeep S. Jawa,[14] Guoqian Jiang,[14] Xia Jing,[7,14] Marcin P. Joachimiak,[8,15] Steven G. Johnson,[9,14,101] Rishikesan Kamaleswaran,[9,15,78] Thomas George Kannampallil,[15,101] Andrew S. Kanter,[15,52] Ramakanth Kavuluru,[9,13,14] Kamil Khanipov,[8,14] Hadi Kharrazi,[9,14] Dongkyu Kim,[15,52] Boyd M. Knosp,[8,15] Arunkumar Krishnan,[9] Tahsin Kurc,[9,15] Albert M. Lai,[101] Christophe G. Lambert,[52,101] Michael Larionov,[14] Stephen B. Lee,[1,14] Michael D. Lesh,[9] Olivier Lichtarge,[14] John Liu,[9] Sijia Liu,[8,9,101] Hongfang Liu,[9,15] Johanna J. Loomba,[1,15,78,101] Sandeep K. Mallipattu,[9,14,15] Chaitanya K. Mamillapalli,[14] Christopher E. Mason,[15] Jomol P. Mathew,[8,15,52] James C. McClay,[101] Julie A. McMurry,[1,4,7,9,13,14,78] Paras P. Mehta,[14] Ofer Mendelevitch,[9] Stephane Meystre,[8,14,15] Richard A. Moffitt,[9,13,15] Jason H. Moore,[8,9] Hiroki Morizono,[13,14,15,52] Christopher J. Mungall,[15,52] Monica C. Munoz-Torres,[7,10,78] Andrew J. Neumann,[78] Xia Ning,[14] Jennifer E. Nyland,[13,14] Lisa O'Keefe,[78] Anna O'Malley,[78] Shawn T. O'Neil,[78] Jihad S. Obeid,[10,14,15] Elizabeth L. Ogburn,[13] Jimmy Phuong,[9,15,52,100,101] Jose D Posada,[8,15] Prateek Prasanna,[14,52] Fred Prior,[9,14,15] Justin Prosser,[9,78] Amanda Lienau Purnell,[101] Ali Rahnavard,[9,52] Harish Ramadas,[9,52,78] Justin T. Reese,[9,10] Jennifer L. Robinson,[14,100] Daniel L. Rubin,[101] Cody D. Rutherford,[9,101] Eugene M. Sadhu,[8,15] Amit Saha,[9] Mary Morrison Saltz,[15,52,101] Thomas Schaffter,[78] Titus KL Schleyer,[14] Soko Setoguchi,[8,14,15] Nigam H. Shah,[8,14] Noha Sharafeldin,[14] Evan Sholle,[15,52] Jonathan C. Silverstein,[15,52,101] Anthony Solomonides,[101] Julian Solway,[14,101] Jing Su,[101] Vignesh Subbian,[9,52,101] Hyo Jung Tak,[15] Bradley W. Taylor,[9,14] Anne E. Thessen,[14,101] Jason A. Thomas,[15] Umit Topaloglu,[15,52] Deepak R. Unni,[8,9,15,52] Joshua T. Vogelstein,[14] Andréa M. Volz,[7] David A. Williams,[14,15] Kelli M. Wilson,[9,78] Clark B. Xu,[8,9,15] Hua Xu,[9,10,14] Yao Yan,[9,15,52] Elizabeth Zak,[8,15] Lanjing Zhang,[101] Chengda Zhang,[14] Jingyi Zheng,[14]

[1]CREDIT_00000001 (Conceptualization) [4]CREDIT_00000004 (Funding acquisition) [7]CRO_0000007 (Marketing and Communications) [8]CREDIT_00000008 (Resources) [9]CREDIT_00000009 (Software role) [10]CREDIT_00000010 (Supervision role) [13]CREDIT_00000013 (Original draft) [14]CREDIT_00000014 (Review and editing) [15]CRO_0000015 (Data role) [52]CRO_0000052 (Standards role) [78]CRO_0000078 (Infrastructure role) [100]Clinical Use Cases [101]Governance

Open science and team science at an unprecedented scale in clinical informatics!