The National COVID Cohort Collaborative (N3C): A social experiment in collaborative research

Christopher G Chute, MD DrPH
Bloomberg Distinguished Professor of Health Informatics
Professor of Medicine, Public Health, Nursing
Chief Research Information Officer, Johns Hopkins Medicine
Deputy Director, Institute for Clinical and Translational Research
Head, Section of Biomedical Informatics and Data Science

Disclosures

- None

From Practice-based Evidence to Evidence-based Practice

JOHNS HOPKINS MEDICINE

Data

Clinical Warehouse

Registries and Marts

Inference

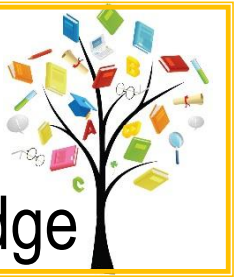
Patient Encounters

Comparable and Consistent

Harmonization

Semantic and Syntactic Standards

Medical Knowledge



Decision Support

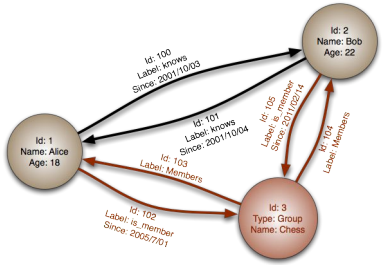
Expert Systems

Clinical Guidelines

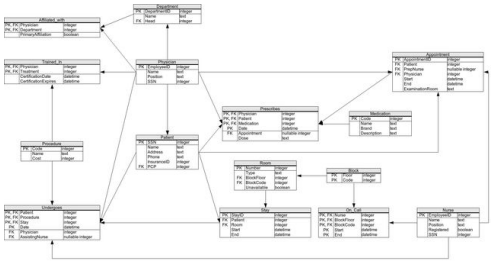
Knowledge Management

Layers of Data Representation
Source, Harmonization, and Analyses

1. Object Layer - Data Transfer, System to System
   a. HL7 FHIR, APIs



2. Integration Layer - Harmonization, Commons
   a. Relational Data



3. Analyses Layer - input for inference (SAS, R, ML)
   a. Flat files (csv/tsv)

National COVID Cohort Collaborative (N3C)

- A **centralized**, secure portal for hosting patient-level COVID clinical data and deploying and evaluating methods and tools for clinicians, researchers, and healthcare

National
COVID
Cohort
Collaborative

- A **partnership** among CTSA program institutions, distributed clinical data networks (e.g. PCORnet, OHDSI, ACT/i2b2, and TriNetX) and many other clinical partners and collaborators
- A partnership between CD2H and NCATS

National Center for Advancing Translational Sciences

National COVID Cohort Collaborative

A program of NIH's National Center for Advancing Translational Sciences

N3C Snapshot

NATIONAL CENTER FOR DATA TO HEALTH

Over 13B Rows of Data

COVID-19 Positive Patients	Total Patients	Sites	Rows of Data
4,303,428	11,924,180	70	13.4b

Procedures	Lab Results	Drug Exposures	Visits	Observations
634.0m	6.4b	2.0b	659.2m	1.2b

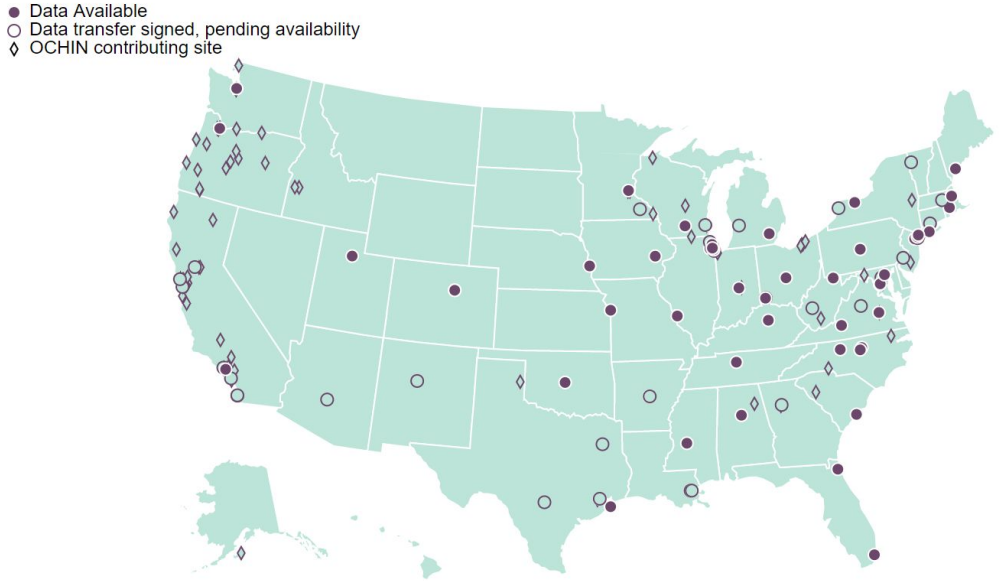
National COVID Cohort Collaborative

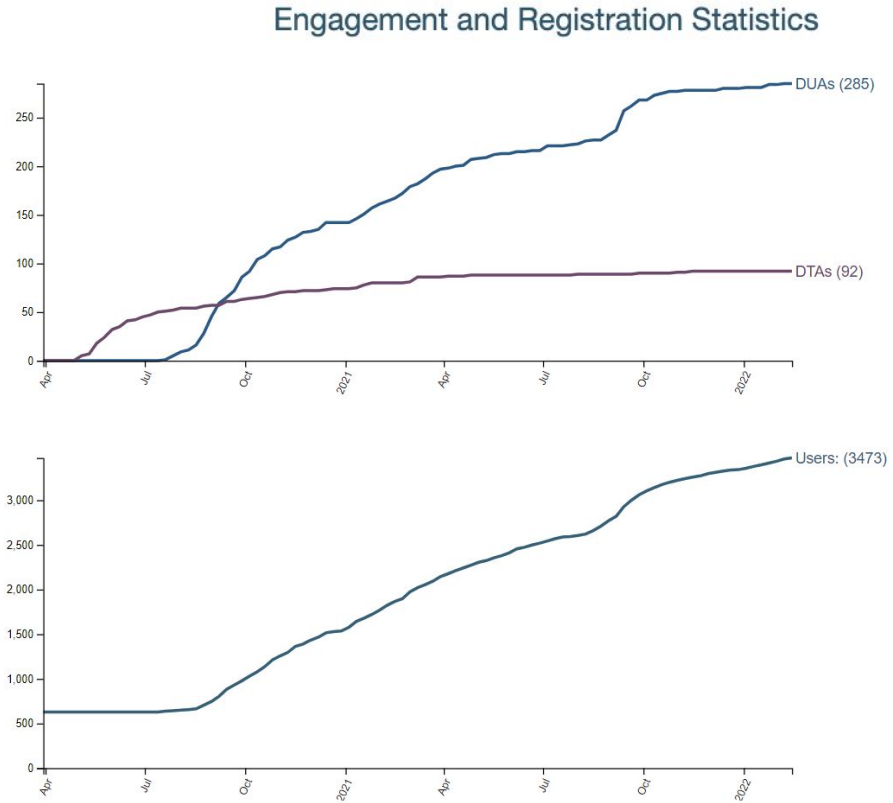
A program of NIH's National Center for Advancing Translational Sciences

N3C Dashboard

covid.cd2h.org/dashboard

NATIONAL CENTER FOR DATA TO HEALTH





covid.cd2h.org/teams

31 Domain teams!

Responsibility

- Largest publicly accessible repository of longitudinal EHR data ever assembled
- Demands unprecedented level of governance
- Enables broad access, team science, and multidisciplinary collaboration

National COVID Cohort Collaborative

A program of NIH's National Center for Advancing Translational Sciences

N3C Community Guiding Principles

NATIONAL CENTER FOR DATA TO HEALTH

Partnership: N3C community members are trusted partners committed to honoring the N3C community guiding principles and Code of Conduct.

Inclusivity: N3C is open to any organization that wishes to contribute data, code, and ideas, as well as anyone who registers to use N3C data to conduct COVID-19 related research, including citizen/community scientists.

Transparency: Open processes and reproducible research is the hallmark of N3C and good scientific practice. Access to data is project-based and focused on COVID-19 research questions. Descriptions of projects are posted publicly and are searchable to promote collaborations.

Reciprocity: Contributions are acknowledged and results from analyses, including provenance and attribution, are expected to be shared with the N3C community.

Accountability: N3C community members take responsibility for their activities and hold each other accountable for achieving the N3C objectives and acting through good scientific practices.

Security: All activities are conducted in a secure, controlled access cloud-based environment, and are recorded for auditing and attribution purposes.

Mutual Respect: Communications should be professional, concise, clear and relevant. Follow proper communication etiquette. Avoid excessive conflict, unprofessional arguments, ad hominem attacks, and/or ridicule over chat and in messaging.

National COVID Cohort Collaborative

Governance & sIRB Overview

Single/Central IRB (sIRB)

- Johns Hopkins serving as central IRB
- Smart IRB makes it easy - all CTSAs are already members, so if you're willing to rely on sIRB, the paperwork is basically complete
- Not required - if you want to do the work locally, you can do so

Who to contact about reliance or local filing

Tricia Francis pfranci4@jhu.edu

JOHNS HOPKINS UNIVERSITY

SMART IRB

National COVID Cohort Collaborative

N3C Governance

Community-centered governance

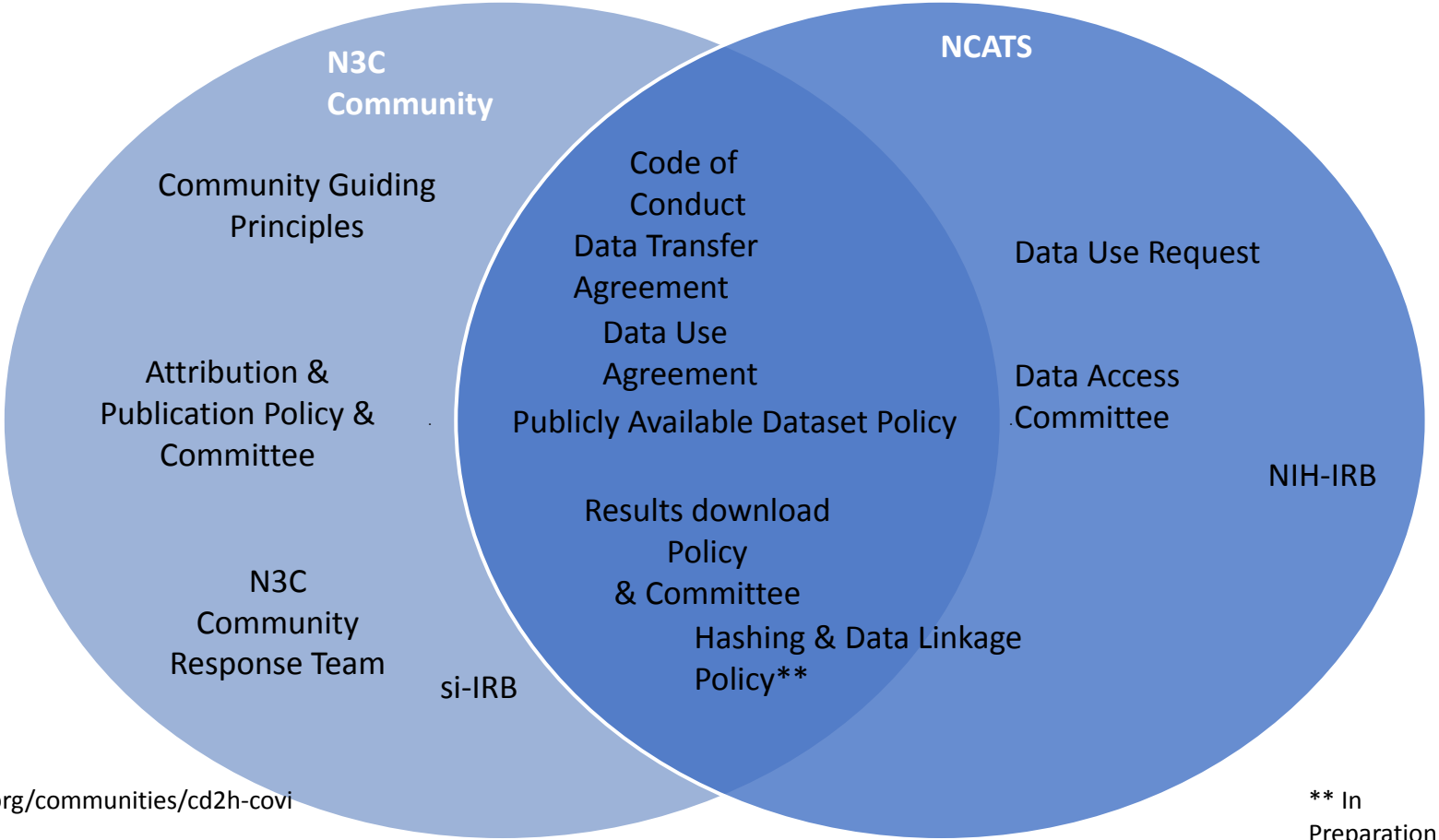
- **N3C Community**: Community norms, team science, workstreams, joined publications
- **NCATS**: N3C Enclave, contribution of Limited Datasets, oversight of data usage

Data Partnership & Governance Workstream

- Establish the principles, policies, and best practices to support the N3C
- Create the functions and roles to implement the policies and strategies
- Design the infrastructure and procedures for controls and audits

National COVID Cohort Collaborative

Shared Governance

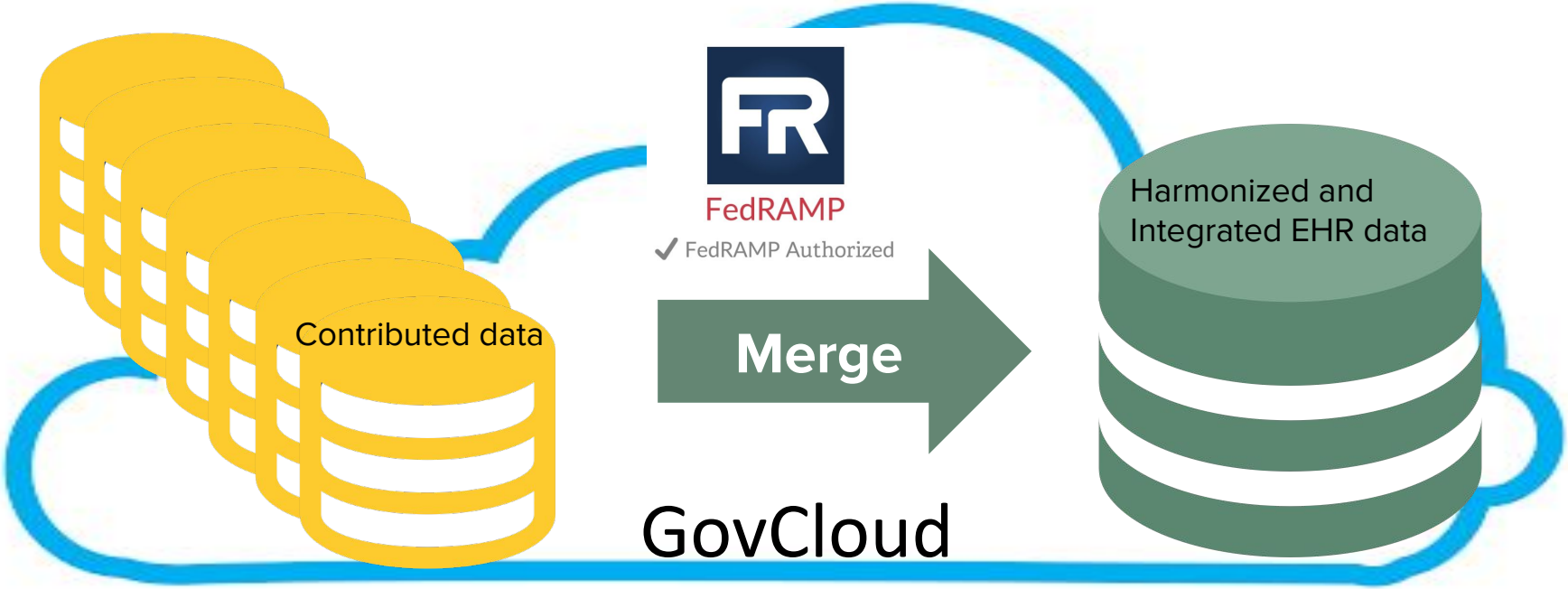


https://zenodo.org/communities/cd2h-covid/

** In Preparation

National COVID Cohort Collaborative

N3C Data Harmonization



Versioned data from all sources is combined into a target model (OMOP)

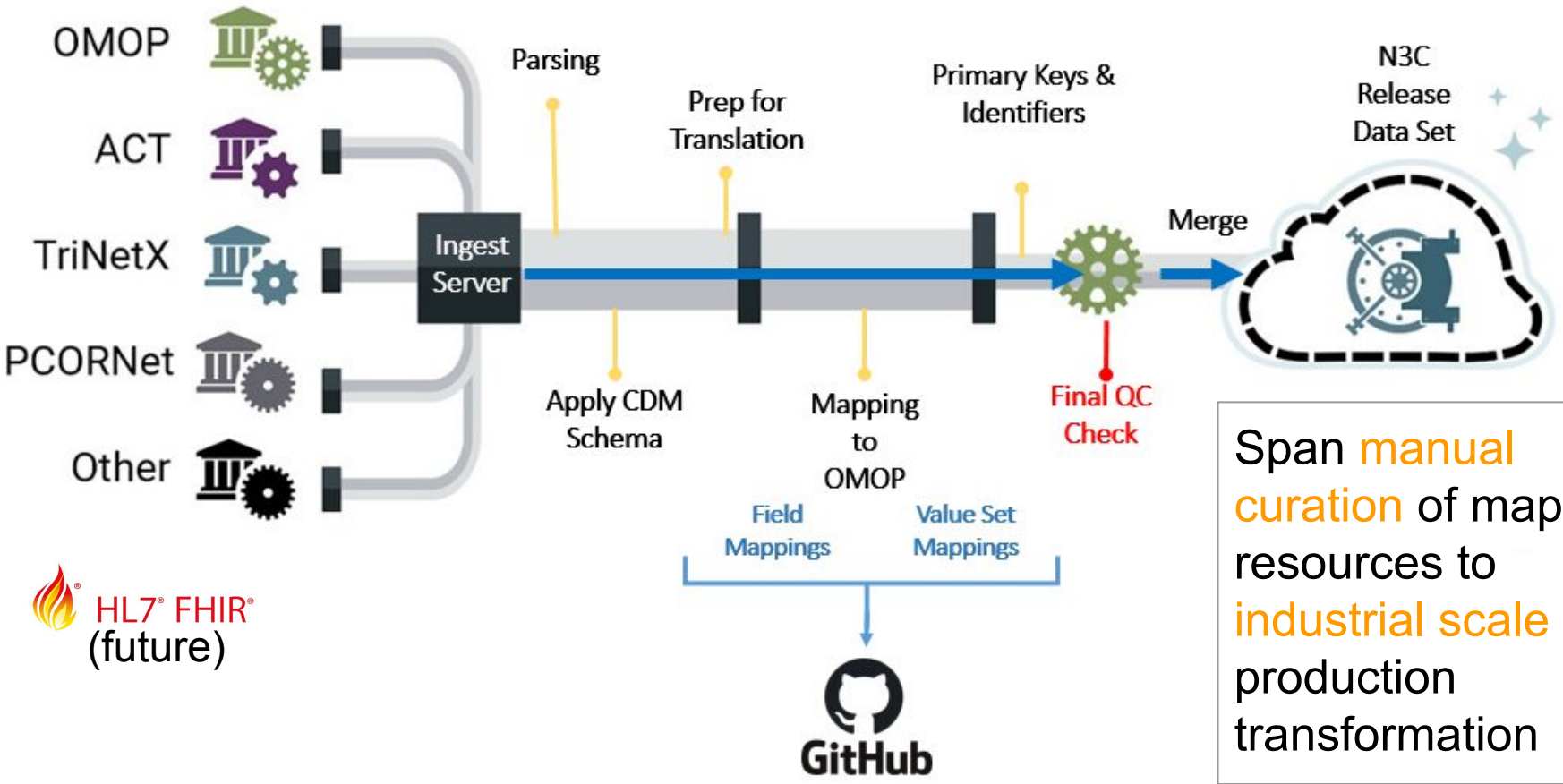
Leveraging Common Data Models

- These four data models are commonly used by academic medical centers throughout the US.
- CDMs are used to store EHR data in a consistent way.
- Sites participating in N3C may send data in one of these four formats—the idea is to make it as convenient as possible for sites to submit.
- Common data models *also* allow us to write a consistent computable phenotype that can be run with few local changes at sites with one or more of these data models.

National COVID Cohort Collaborative
A program of NIH's National Center for Advancing Translational Sciences

N3C Data Ingestion & Harmonization Pipeline

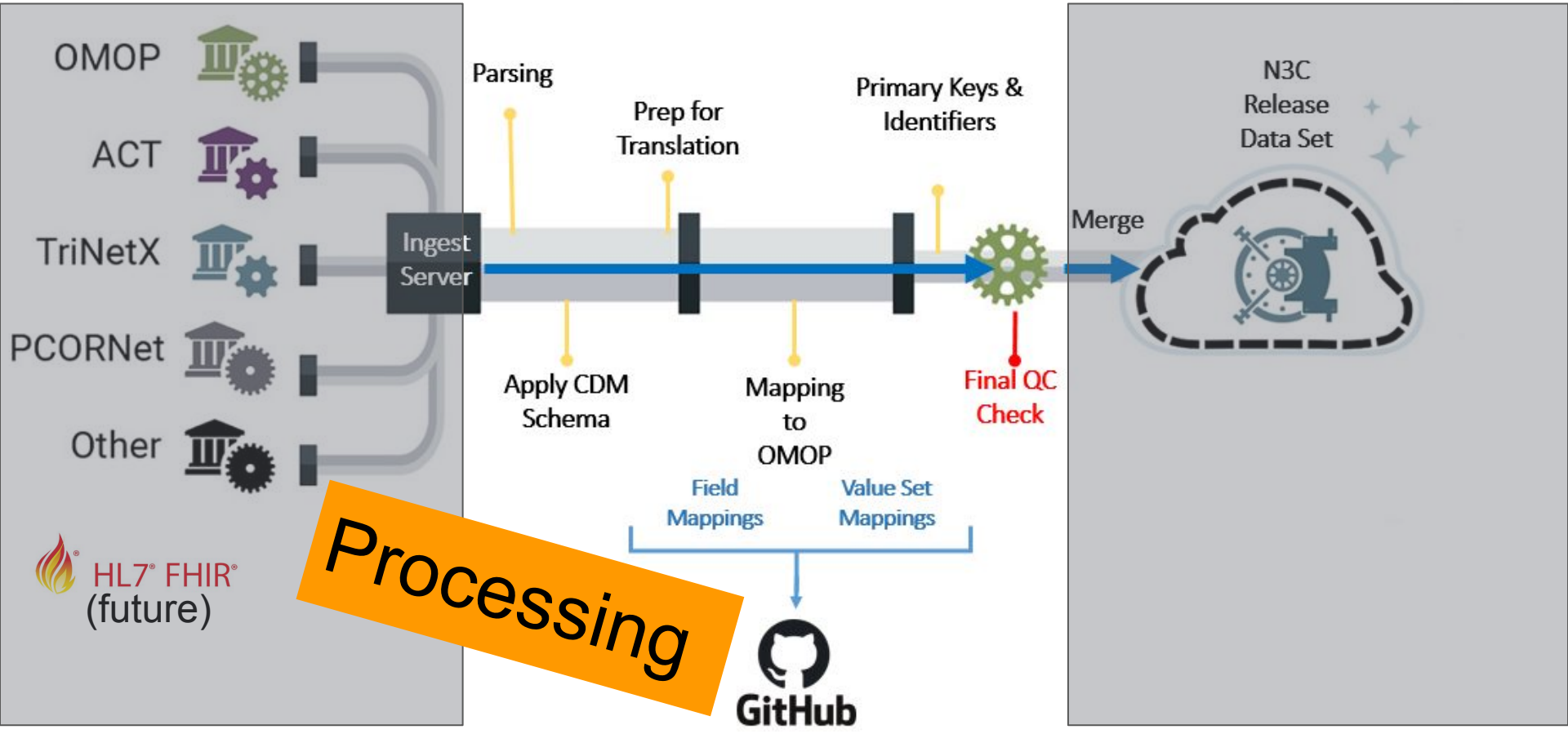
NATIONAL CENTER FOR DATA TO HEALTH



Span manual curation of mapping resources to industrial scale production transformation

National COVID Cohort Collaborative

N3C Data Ingestion & Harmonization Pipeline



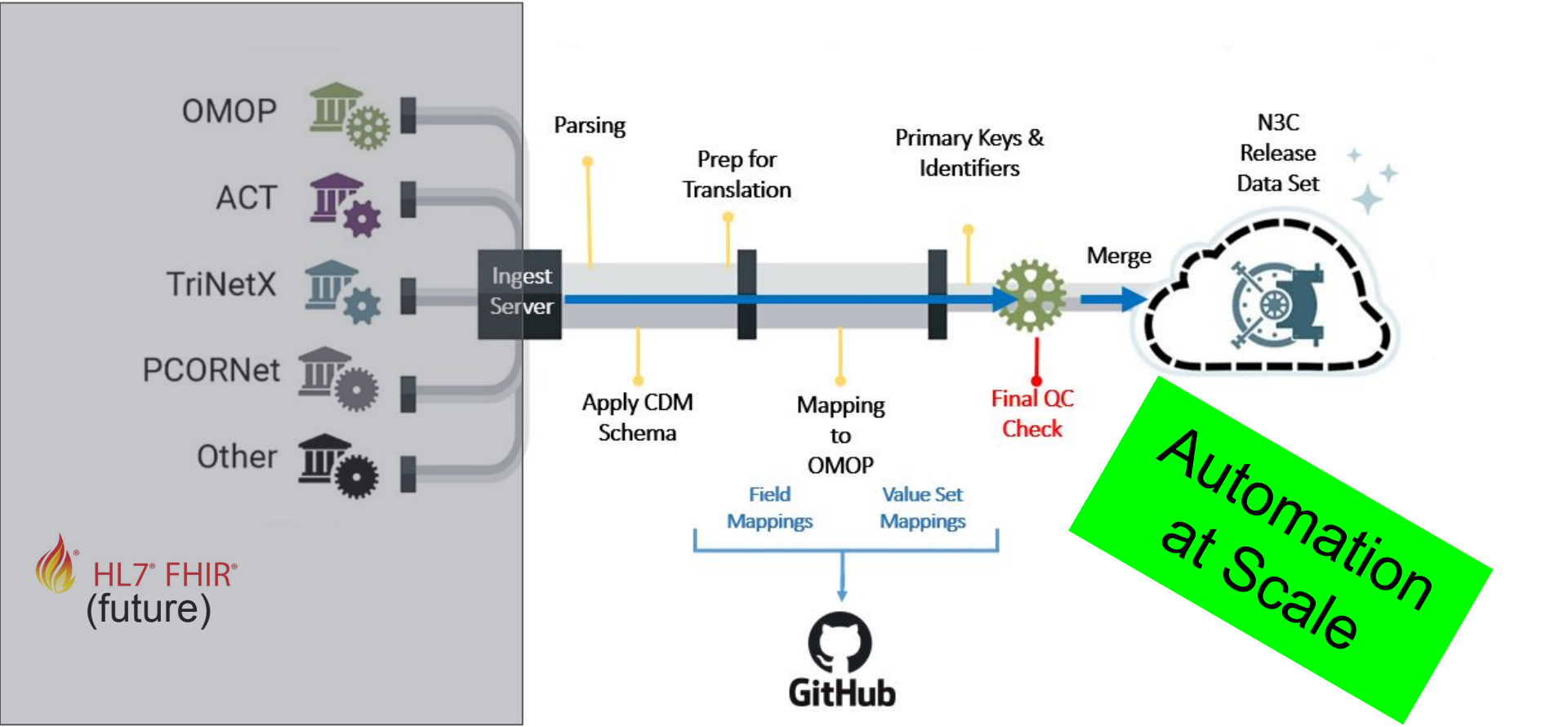
National COVID Cohort Collaborative

Human mediated mapping and validation

- About 2M structure, terms, codes mapped between common data models
- Many mappings were leveraged or expanded from pre-existing work
- All were validated for this project
- Validation required human curation and sampled cross-checking
- This process parallels much work already done for CRDC by CCDH and the creation of CRDC-H versions

National COVID Cohort Collaborative

N3C Data Ingestion & Harmonization Pipeline

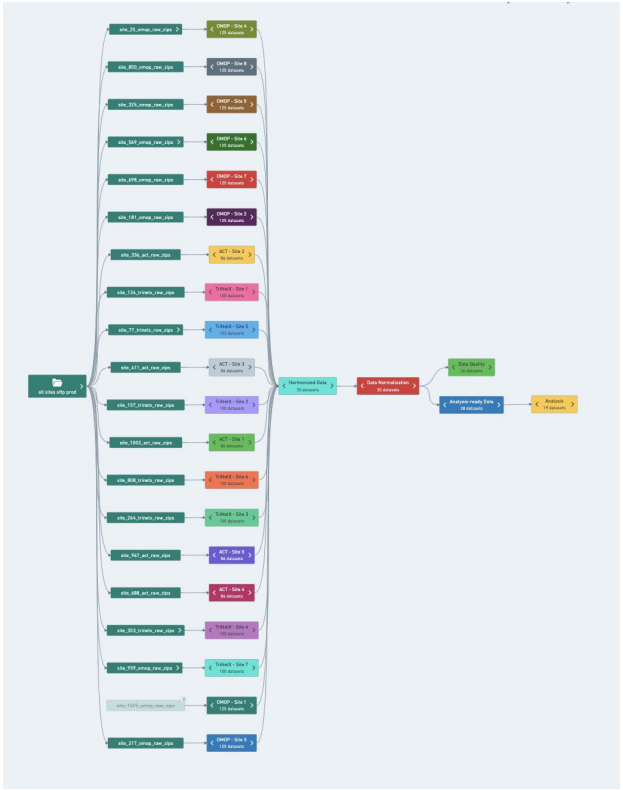


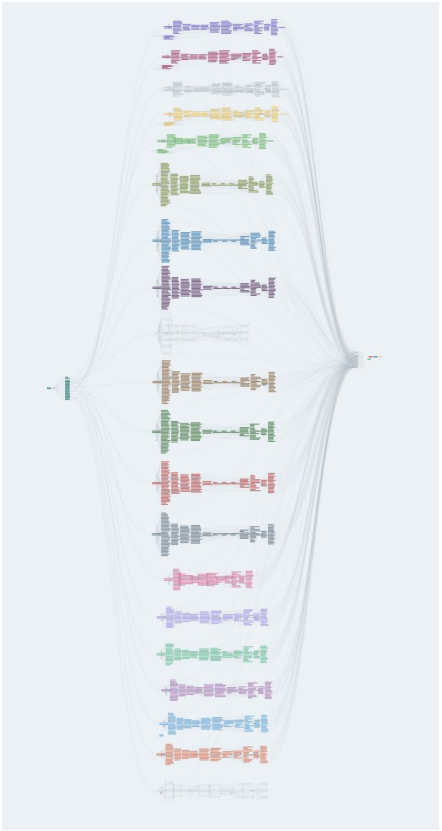
National COVID Cohort Collaborative

A program of NIH's National Center for Advancing Translational Sciences

Each of the 70+ sites has a pipeline with 100+ transformations

NATIONAL CENTER FOR DATA TO HEALTH





The provenance between 5000 transformations across the 70 sites is automatically tracked.

This enables:

- pipeline developers to very quickly identify the root cause of data quality issues
- data pipelines to be refreshed in <20 minutes whenever the source data updates

National COVID Cohort Collaborative

A program of NIH's National Center for Advancing Translational Sciences

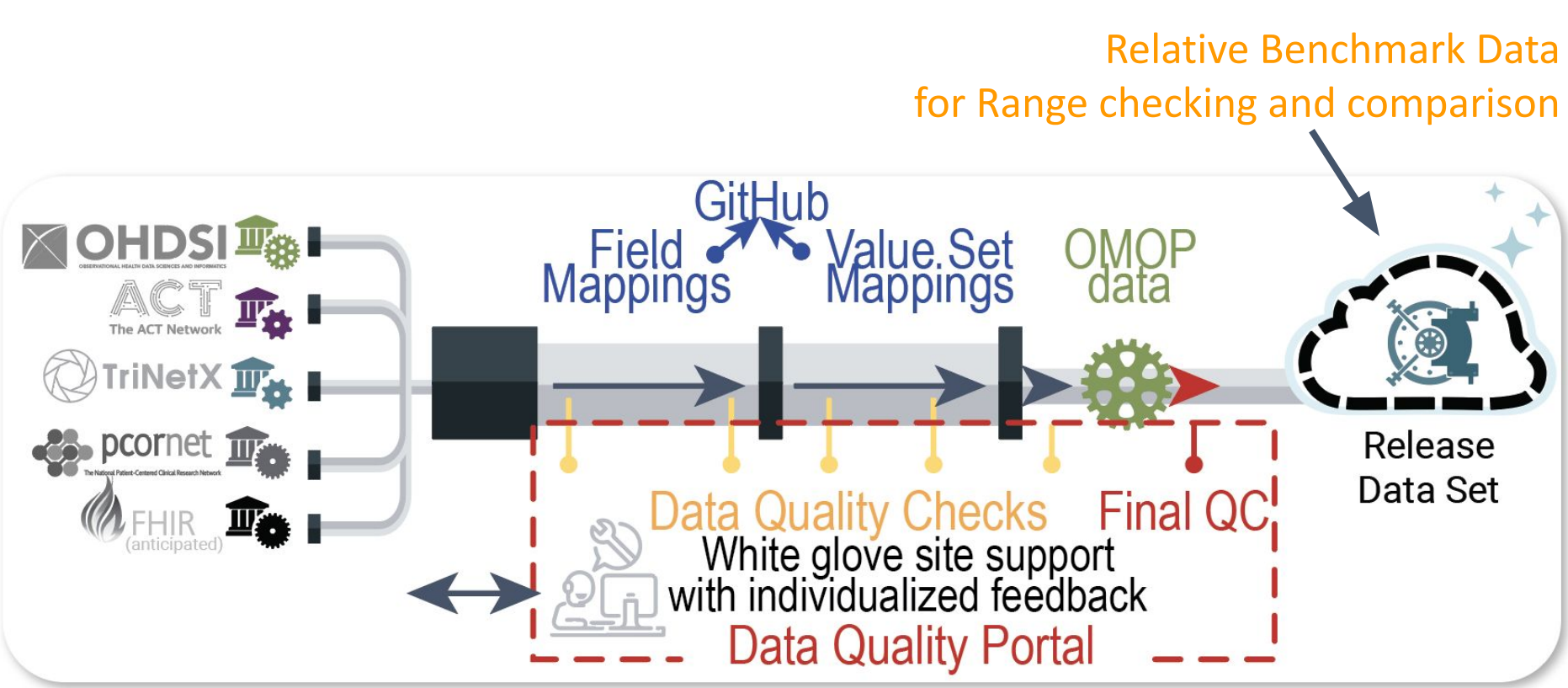
Each site has its own set of data health checks that run each time new data is submitted

NATIONAL CENTER FOR DATA TO HEALTH

- When the CDM mapping pipeline is deployed for a new site, it comes with a set of automated data health checks.
- These run every time the data updates - so that if new data doesn't meet expectations, the pipeline administrators are immediately alerted and can take action

National COVID Cohort Collaborative

Centralized Ingestion

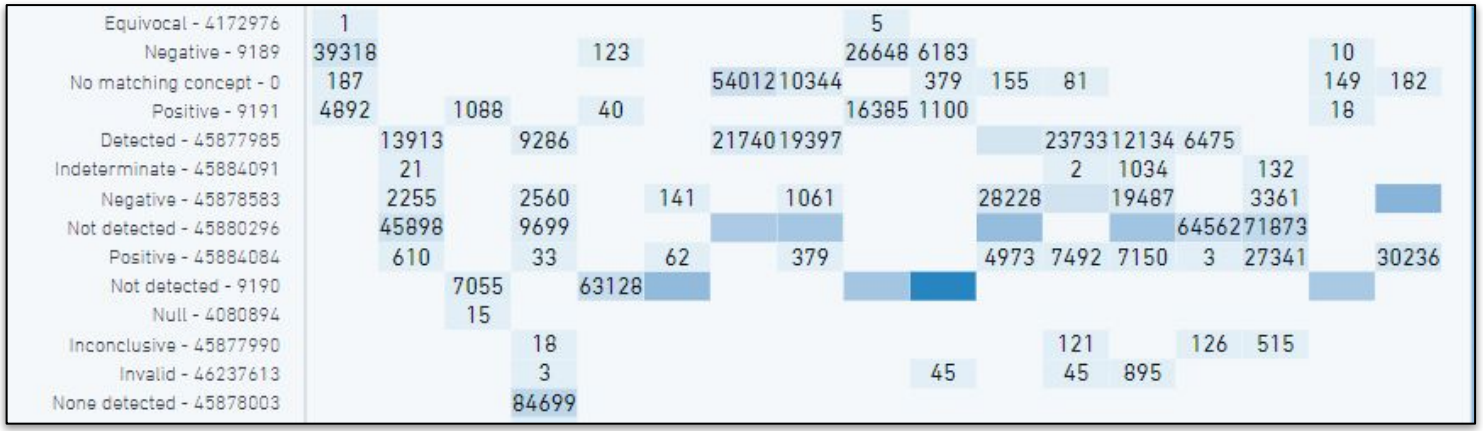


National COVID Cohort Collaborative
A program of NIH's National Center for Advancing Translational Sciences

N3C's DQ Process

NATIONAL CENTER FOR DATA TO HEALTH

It's not all about inpatient visits. N3C data harmonization staff look at a number of data quality metrics for each site, multiple times per week.



Number of COVID tests and proportion of + to -

Average rows of data per patient per table

Conformance to OMOP standards

Plausible demographic distributions

Visits with negative lengths

Harmonizing numeric data

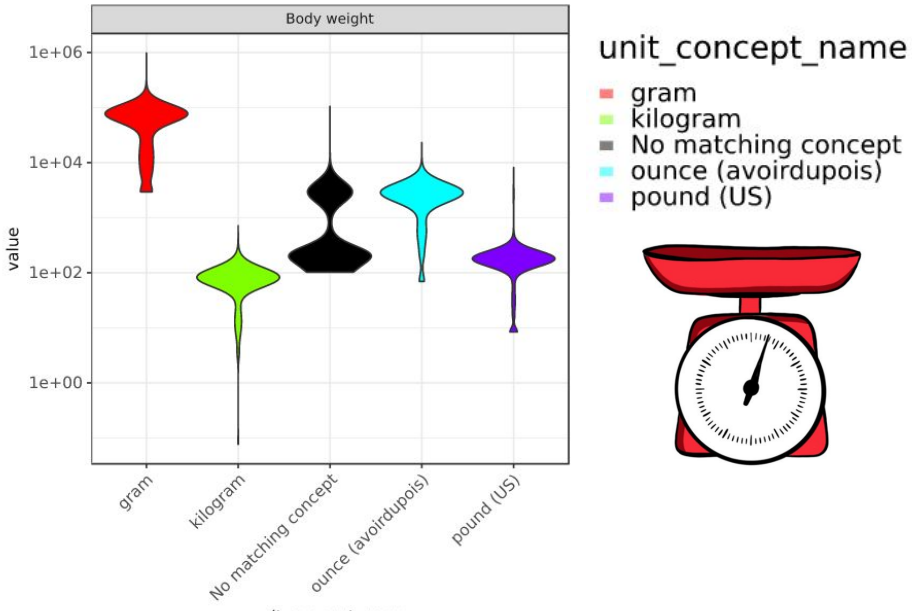
- **Problem:** Different sites provide their data in different units

- **Solution:** Harmonize each to a standard unit

  Kilograms = **Pounds** / 2.20462

  Kilograms = **Ounces** / 35.274

  Kilograms = **Grams** / 1000



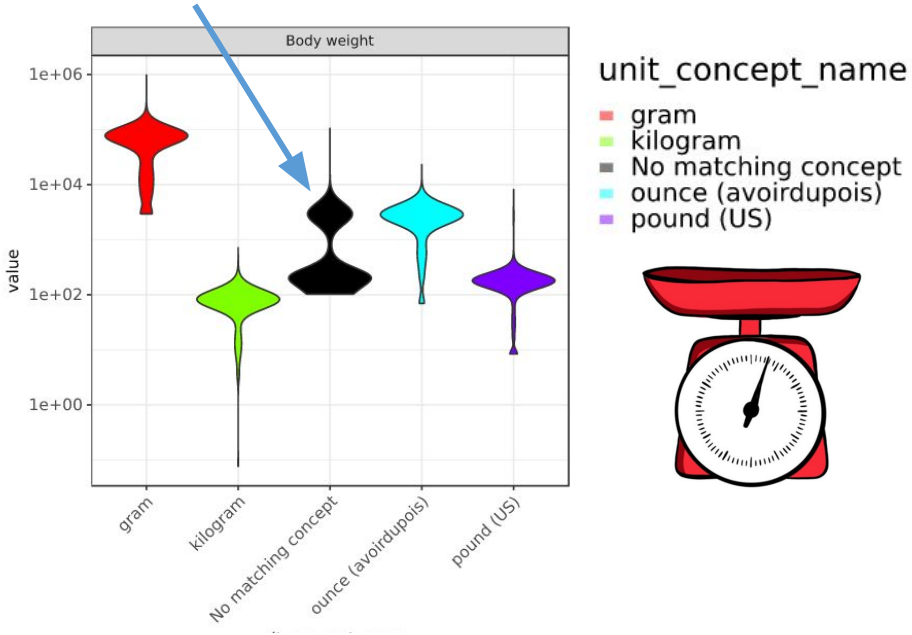
Harmonizing numeric data

- **Problem:** Some units are **missing**
- **Solution 1:** Contact the source
- **Solution 2:** N3C inference engine

  Kilograms = **x** / 2.20462 ?

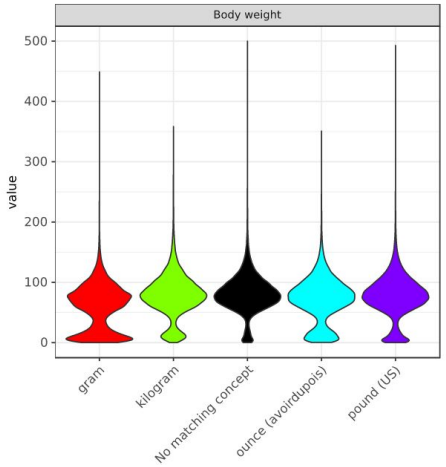
  Kilograms = **x** / 35.274 ?

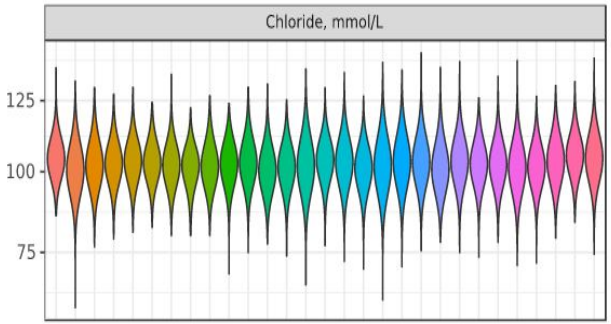
  Kilograms = **x** / 1000 ?



Harmonization progress

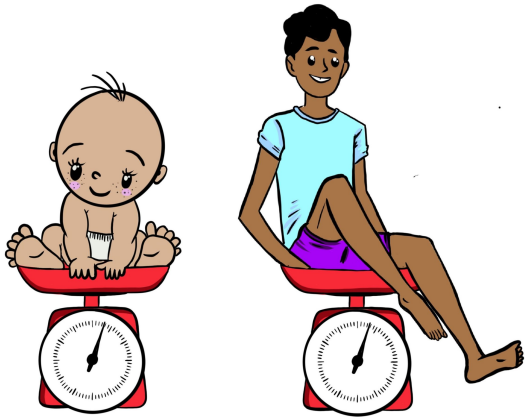
- Harmonized measurements
  - By original unit
  - Across many sites





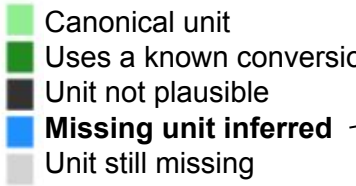
Homogeneity after harmonization

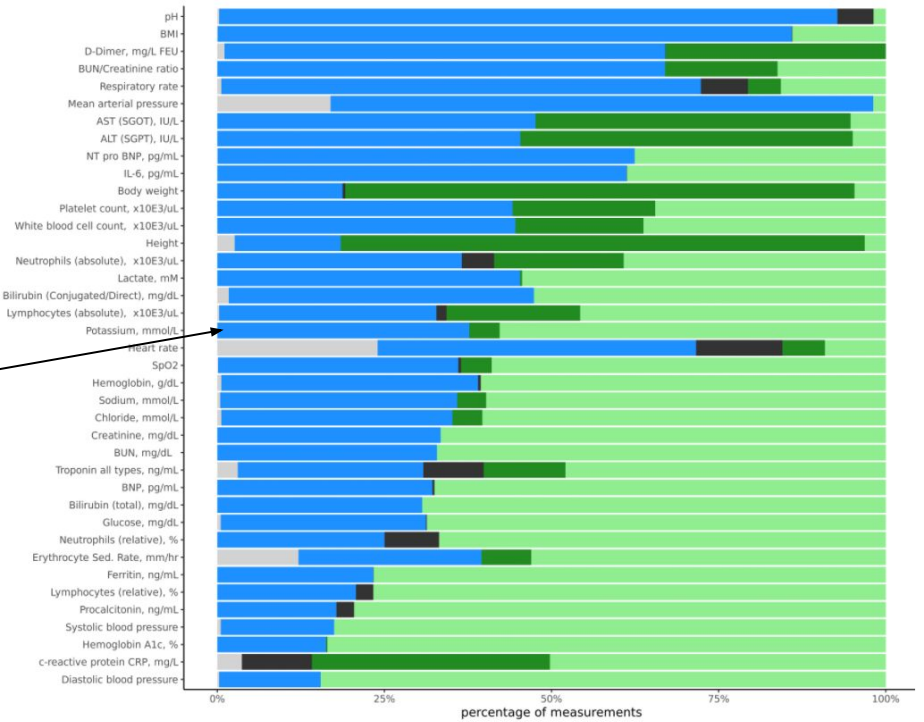
Humans measured in grams do not look the same as humans measured in kilograms!



Unit harmonization progress

- ~2x increase in usable data from our harmonization procedures

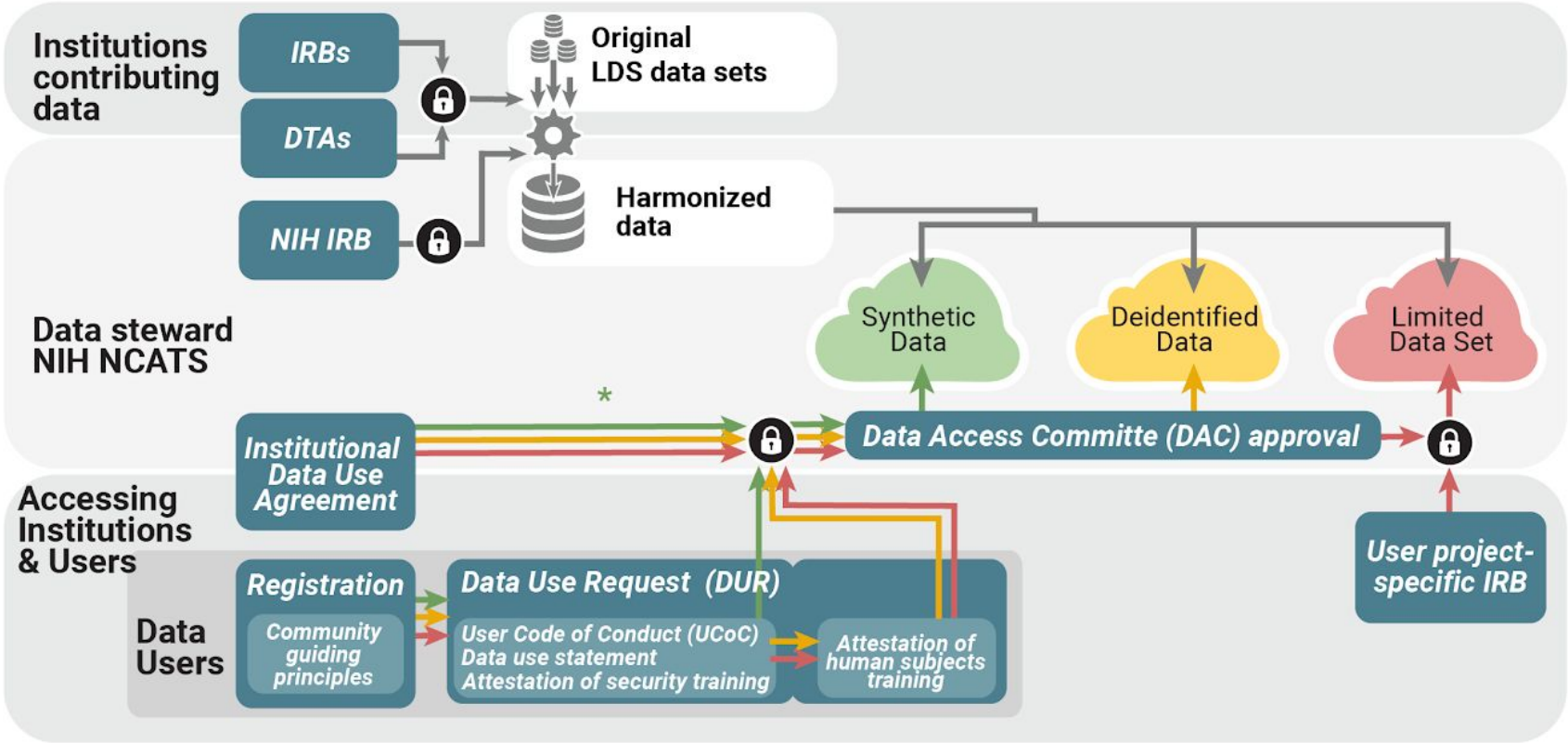




We can rescue a lot of data!

National COVID Cohort Collaborative

N3C: Governance and Access



National COVID Cohort Collaborative

N3C Team Science

Key functions can nucleate projects:

- Education & training
- Biostatistics
- Study design
- Evaluation
- Informatics
- Clinical expertise
- Innovation & commercialization
- Community & partnerships

N3C Domain Team Expertise:

- Enclave technology
- Data model (OMOP)
- Terminologies
- Data quality
- Codesets, variables, phenotype
- Using/parsing N3C data
- Workflows, methods, algorithms

covid.cd2h.org/Domain_Teams

National COVID Cohort Collaborative

A program of NIH's National Center for Advancing Translational Sciences

Domain Teams

NATIONAL CENTER FOR DATA TO HEALTH

Ask N3C

N3C Domain Teams

N3C Domain Teams enable researchers with shared interests analyze data within the N3C Data Enclave and collaborate more efficiently in a team science environment. These teams provide an opportunity to collect pilot data for grant submissions, train algorithms on larger datasets, inform clinical trial design, learn how to use tools for large scale COVID-19 data, and validate results. Domain Teams are enabled by Slack channels for discussion, meetings, and document management and are supported by N3C workstreams. N3C encourages researchers of all levels to join a Domain Team that represents their interests, or to suggest new clinical areas to explore. A Domain Team can submit one or more research projects, but collaboration is encouraged for similar concepts.

Types of Domain Teams

Multi-discipline **Clinical Domain Teams** focus on clinical questions surrounding COVID-19's impact on health conditions and consist of clinical and subject matter experts, statisticians, informaticists, and machine learning specialists. **Cross-Cutting Domain Teams** have a varied focus that applies to multiple domains.

Start a Domain Team | Domain Team Support

To get started with N3C and gain access to the N3C Data Enclave, please view the Researcher Essentials webpage.

Domain Team Testimonial

National COVID Cohort Collaborative

A program of NIH's National Center for Advancing Translational Sciences

Highly Active Domain Team Calendar

NATIONAL CENTER FOR DATA TO HEALTH

National COVID Cohort Collaborative

A program of NIH's National Center for Advancing Translational Sciences

Attribution and Publication

NATIONAL CENTER FOR DATA TO HEALTH

N3C Publication Committee was organized to:

(1) Facilitate inclusive and appropriate attribution
(2) Troubleshoot authorship questions
(3) Track the output of N3C community
(4) Promote awareness of disseminated information

Authorship & Attribution

- Masthead (name in byline)
- Block authorship ('N3C Consortium' in byline)
- Acknowledgement

N3C ICMJE Requirements

Yes to 1, 2, and 3
AUTHOR

Less
CONTRIBUTOR

1. Conception & Design
or data acquisition
or analysis & interpretation

↓

2. Drafting manuscript
or revising

↓

3. Final manuscript approval

National COVID Cohort Collaborative
A program of NIH's National Center for Advancing Translational Sciences

Impact to Date - Publications

NATIONAL CENTER FOR DATA TO HEALTH

During 15 months of N3C operation:

Pubmed cited, peer-reviewed:	16
Pubmed cited, pre-prints	11
High profile, non-cited	21
Total	48

https://covid.cd2h.org/dashboard/

Impact to Date - Demonstration

- 13.4B rows Patient Data assembled in one place
- Nearly 3,500 scientists have row-level access
- Scientific productivity is unprecedented

- The sky has not fallen…

Observational Data Bridging Gap - Bench to Bedside

- Real-world data is mostly untapped resource
- Society cannot afford randomized clinical trials for all questions we might need to ask

- EHR data has profound limitations
  - Biased collection sources
  - Missing information

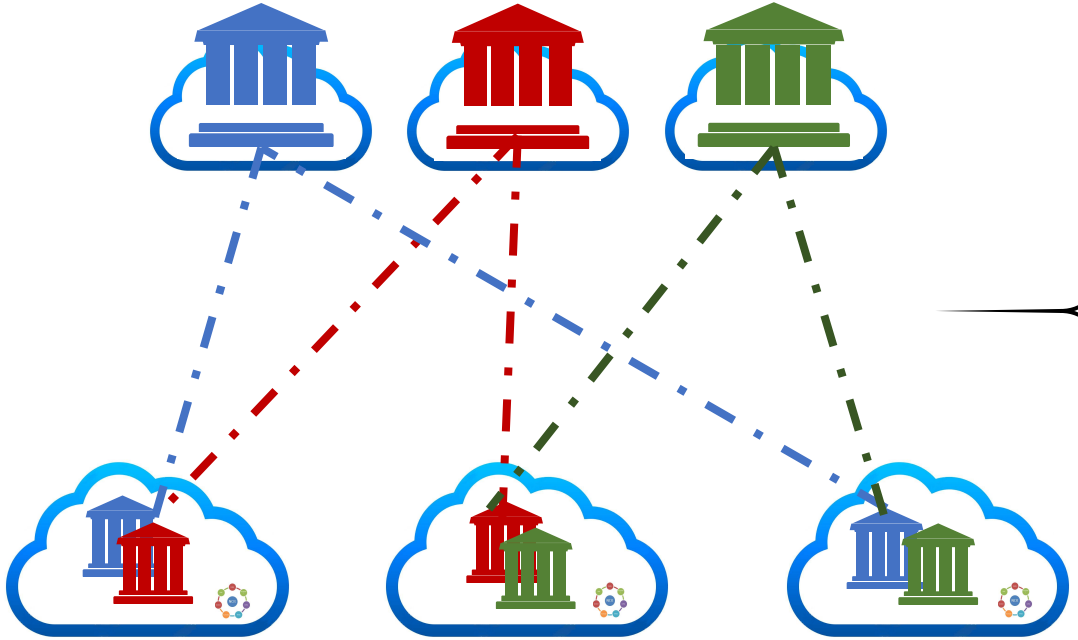
- The “big data” paradigm - enough data informs

The Once and Future "N3C"

- NCATS and NIH highly value what has been built
- Speculations of where it might go follow
- THESE ARE NOT OFFICIAL OR SANCTIONED
- Presented as exemplar
- Something "like this" is likley

Proposal: N3C National ~~COVID~~ Clinical Cohort Collaborative?

Individual Data Contributors



Self-aggregating research networks

- Common Ontology (OMOP)
- Data Cleaning & Harmonisation
- Data Quality
- Data Lineage
- Granular/Purpose-Based Access Controls
- De-identification
- AI/ML Environment
- Code Workbooks
- Analytical Tools

National Center for Advancing Translational Sciences

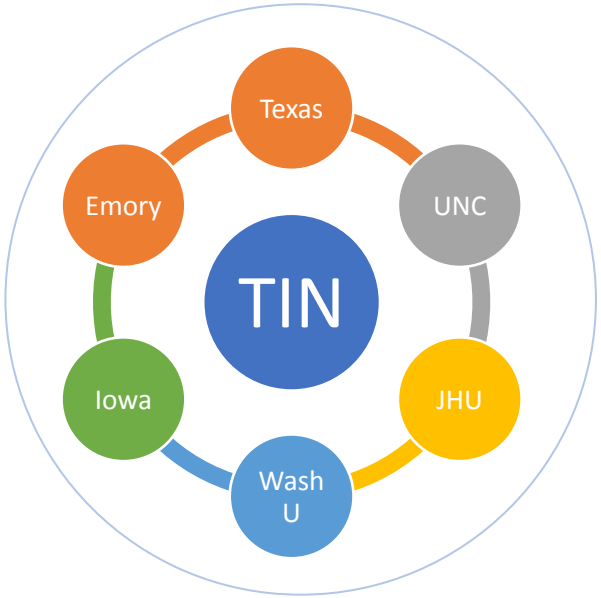
Introducing: National Health Cohort Collaborative (NHC2) [?]

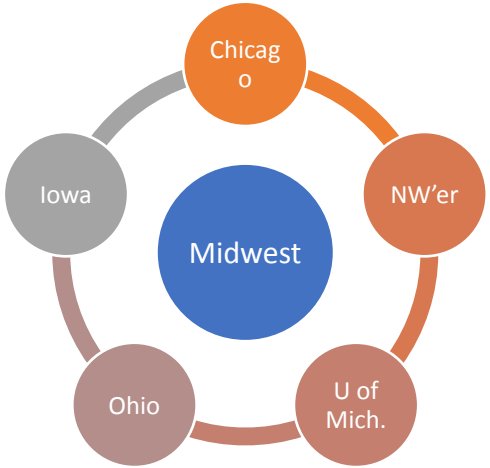
- Build on the principles and success of N3C
- Leverage underlying machinery and infrastructure
  - Maintain “enclave” as core characteristic
- Broaden inclusion criteria
  - Gradually add specific conditions (e.g. rare diseases)
    - Leverage size and scope of CTSA as a Network
  - Evolve toward inclusion of all “longitudinal patients” over all CSTAs
    - May include IDeA CTRs
    - Consider CTSAs engaging state Health Information Exchanges
- Demonstrate synergy between local curation and centralized QA/Benchmarking
- Foster rapid observational biomedical research at scale

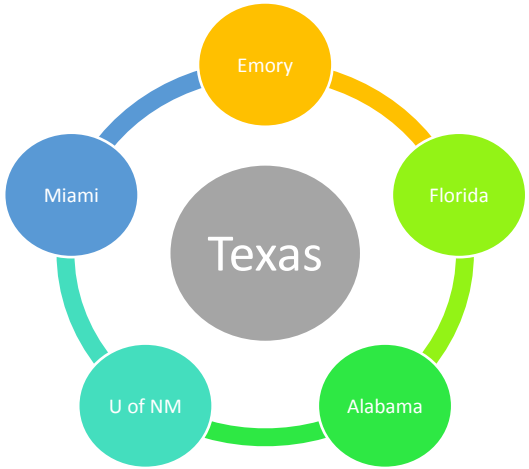
Motivations and Goals of N3C vs NH2C

- N3C was created during a pandemic
  - Collaborations gave benefit of doubt in public health emergency
  - Goal was to bring as many minds as possible to the data
    - Accelerate discovery and understanding urgently
  - Build most-inclusive analyst community practical
    - Data Access Committee exclusively NIH
  - Contributing sites had binary decision: All-in or Out
- NH3C likely to operate under “normal” science-culture rules
  - No emergency or perceived urgency
  - Goal will be to centralize as much harmonized data as practical and useful
  - Primary analyst community likely to arise among data contributors
    - Data Access Committee must be led by contributing organizations
  - Contributing sites will expect agency over contributed data
    - Options: All-in, Domain by domain, Study by study, Out

CTSAs can create self aggregating research networks







QA/QI

Harmonization

Security

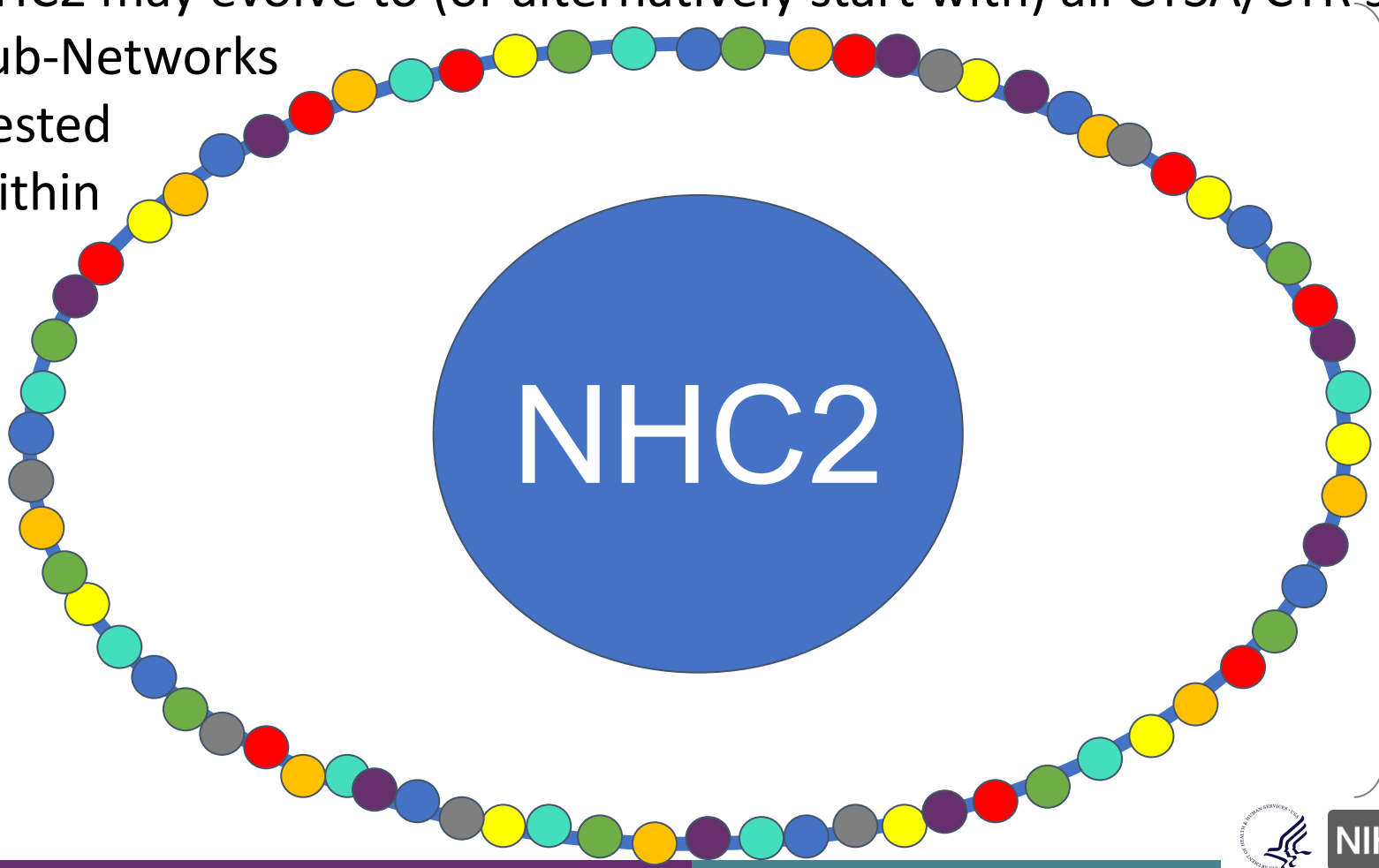
IAAS
PAAS
SAAS

Analytics

National Center for Advancing Translational Sciences

NHC2 may evolve to (or alternatively start with) all CTSA/CTR sites

Sub-Networks nested within



- QA/QI
- Harmonization
- Security
- IAAS
  PAAS
  SAAS
- Analytics

Thank you!

ACCEPTED MANUSCRIPT

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** ▾

OPEN ACCESS

bit.ly/n3c-methods-jamia

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] Linqiang 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

Title

- Text