5.3 Clinical Research Informatics (2/2)

What is Biomedical and Health Informatics? - http://informatics.health/

William Hersh, MD, FACMI, FAMIA, FIAHSI

Solutions

- Integrated data repositories (IDRs)
  - Definitions
  - Growing best practices of extracting EHR data
  - Emerging common data models
  - Exemplars and examples
- Other innovations and lessons learned

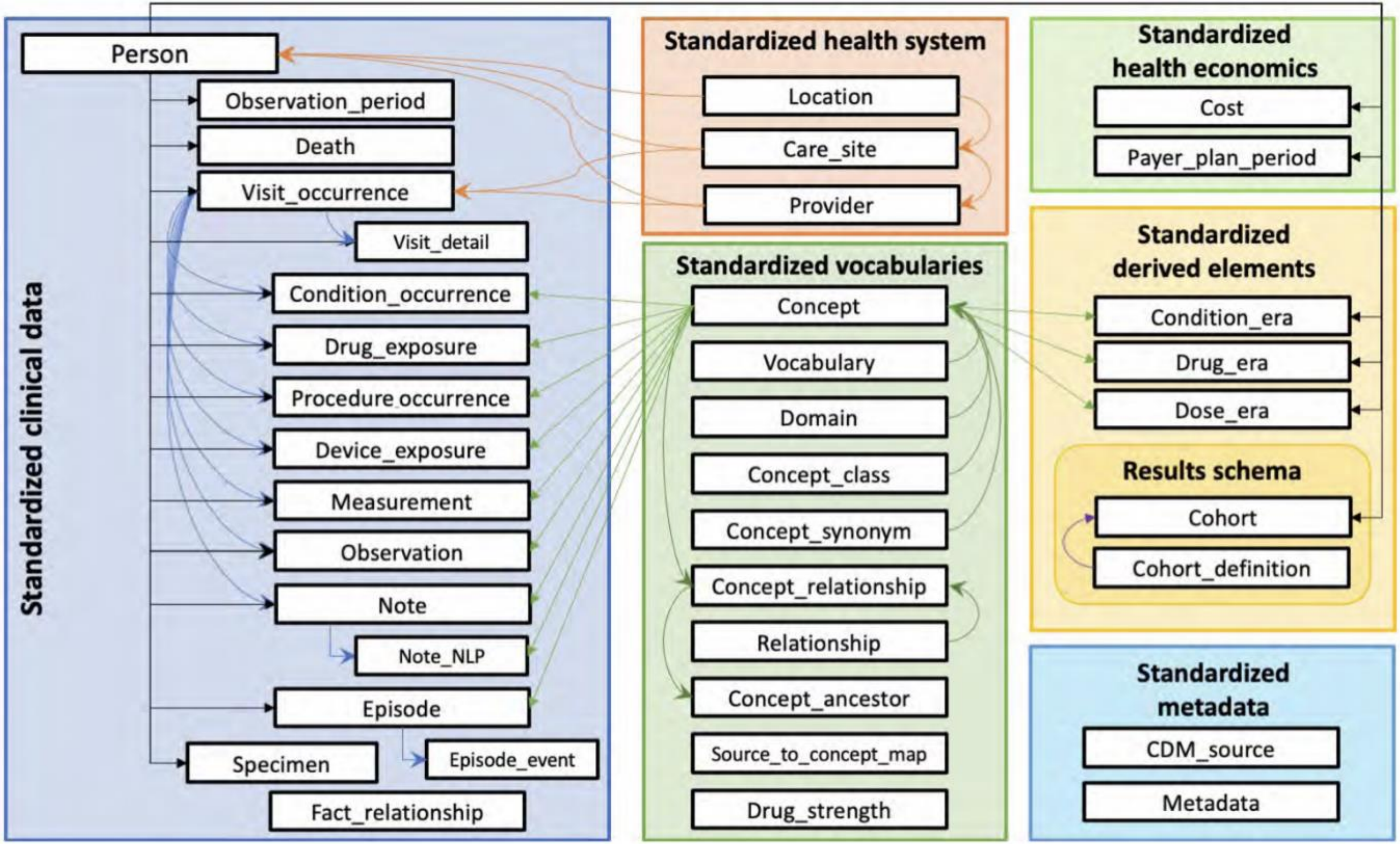
Integrated data repositories (IDRs)

- Combine clinical and research data allowing research and other secondary uses of clinical data (Lin, 2024)
- Techniques such as propensity scores allow adjustment for confounders in observational, e.g., EHR data (Haukoos, 2015)
- Important best practices for
  - Data and service (Campion, 2020)
  - Selection of data models (Schneeweiss, 2020)
  - Architectural choices, e.g., centralized vs. federated data (Li, 2024)

Observational Health Data Sciences and Informatics (OHDSI)

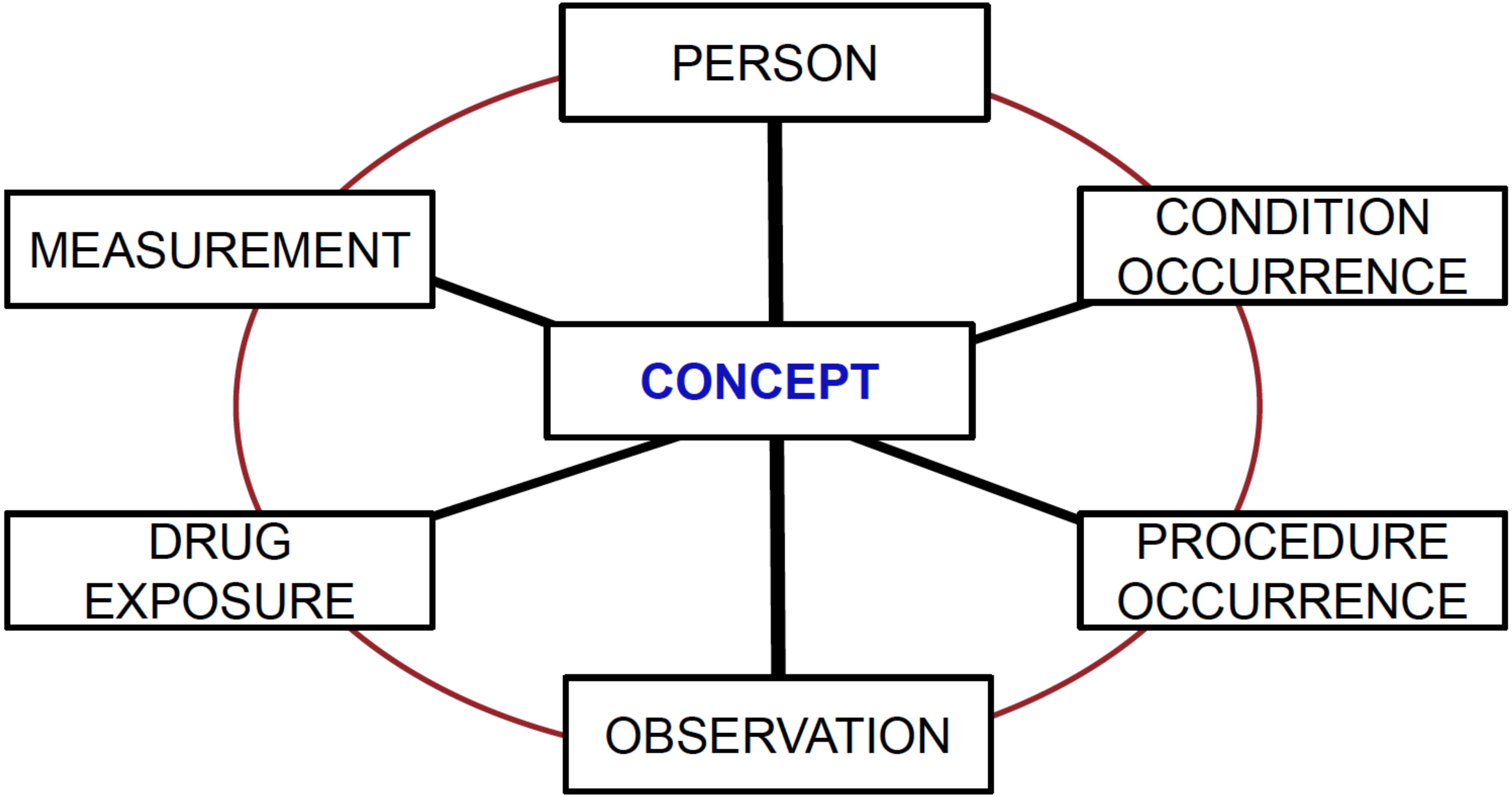
- (Hripcsak, 2015; Hripcsak, 2016; Book)
- Growing global research community of 4000+ collaborators from 83 countries
- Federated database of 331 data sources with 2.1 billion patient records across 34 countries (Reich, 2024)
- All adhere to common data model based on Observational Medical Outcomes Partnership (OMOP)
- OHDSI Standardized Vocabularies comprise over 10 million concepts from 136 vocabularies (Reich, 2024)
- Software tools allow query of data at each site

OMOP Common Data Model (Reich, 2024)

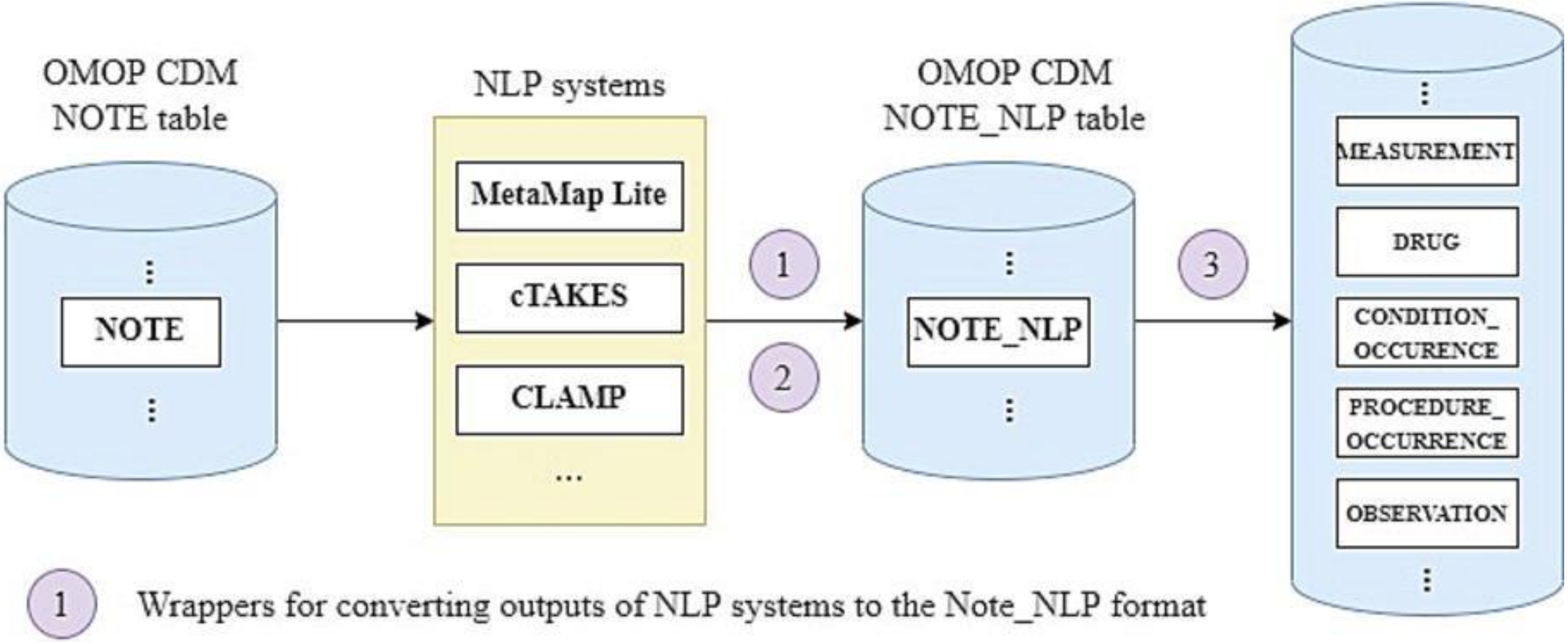


WhatIs5.3

OMOP concept tables



ODHSI also includes clinical notes (Keloth, 2023)

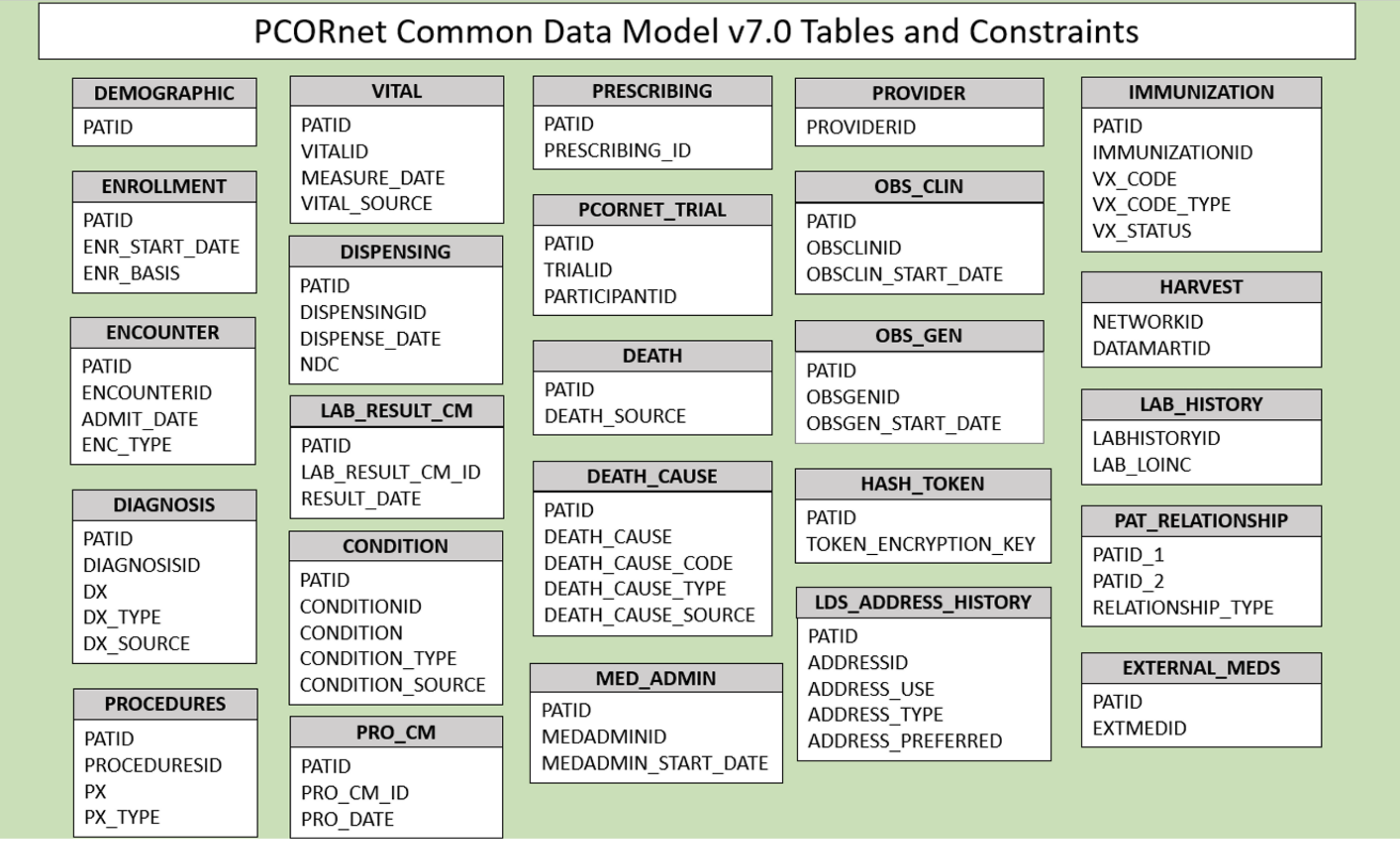


1 Wrappers for converting outputs of NLP systems to the Note_NLP format

2 Mapping concepts to OHDSI vocabulary using Ananke

3 SQL scripts for transferring data from Note_NLP to clinical tables

Common data models – PCORnet



PCORnet Common Data Model v7.0 Tables and Constraints

DEMOGRAPHIC
PATID

ENROLLMENT
PATID
ENR_START_DATE
ENR_BASIS

ENCOUNTER
PATID
ENCOUNTERID
ADMIT_DATE
ENC_TYPE

DIAGNOSIS
PATID
DIAGNOSISID
DX
DX_TYPE
DX_SOURCE

PROCEDURES
PATID
PROCEDURESID
PX
PX_TYPE

VITAL
PATID
VITALID
MEASURE_DATE
VITAL_SOURCE

DISPENSING
PATID
DISPENSINGID
DISPENSE_DATE
NDC

LAB_RESULT_CM
PATID
LAB_RESULT_CM_ID
RESULT_DATE

CONDITION
PATID
CONDITIONID
CONDITION
CONDITION_TYPE
CONDITION_SOURCE

PRO_CM
PATID
PRO_CM_ID
PRO_DATE

PRESCRIBING
PATID
PRESCRIBING_ID

PCORNET_TRIAL
PATID
TRIALID
PARTICIPANTID

DEATH
PATID
DEATH_SOURCE

DEATH_CAUSE
PATID
DEATH_CAUSE
DEATH_CAUSE_CODE
DEATH_CAUSE_TYPE
DEATH_CAUSE_SOURCE

MED_ADMIN
PATID
MEDADMINID
MEDADMIN_START_DATE

PROVIDER
PROVIDERID

OBS_CLIN
PATID
OBSCLINID
OBSCLIN_START_DATE

OBS_GEN
PATID
OBSGENID
OBSGEN_START_DATE

HASH_TOKEN
PATID
TOKEN_ENCRYPTION_KEY

LDS_ADDRESS_HISTORY
PATID
ADDRESSID
ADDRESS_USE
ADDRESS_TYPE
ADDRESS_PREFERRED

IMMUNIZATION
PATID
IMMUNIZATIONID
VX_CODE
VX_CODE_TYPE
VX_STATUS

HARVEST
NETWORKID
DATAMARTID

LAB_HISTORY
LABHISTORYID
LAB_LOINC

PAT_RELATIONSHIP
PATID_1
PATID_2
RELATIONSHIP_TYPE

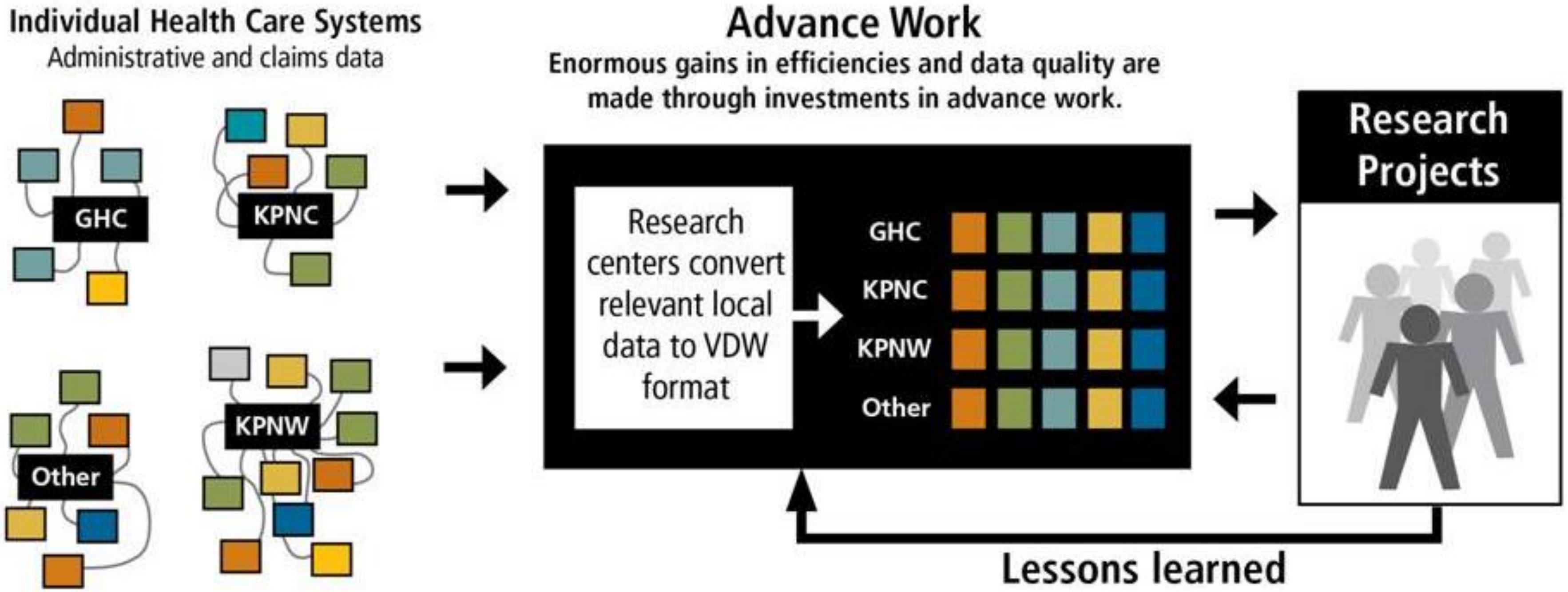
EXTERNAL_MEDS
PATID
EXTMEDID

IDR exemplars and examples

- Health Care Systems Research Network (HCSRN)
- Accessible Research Commons for Health (ARCH), including
  - Informatics for Integrating Biology and the Bedside (i2b2)
  - Shared Health Research Information Network (SHRINE)
- Cosmos Collaborative
- AllOfUs
- National COVID Cohort Collaborative (N3C)
- Merging IDRs together
- Examples of clinical studies
- Lessons learned

Health Care Systems Research Network (HCSRN)

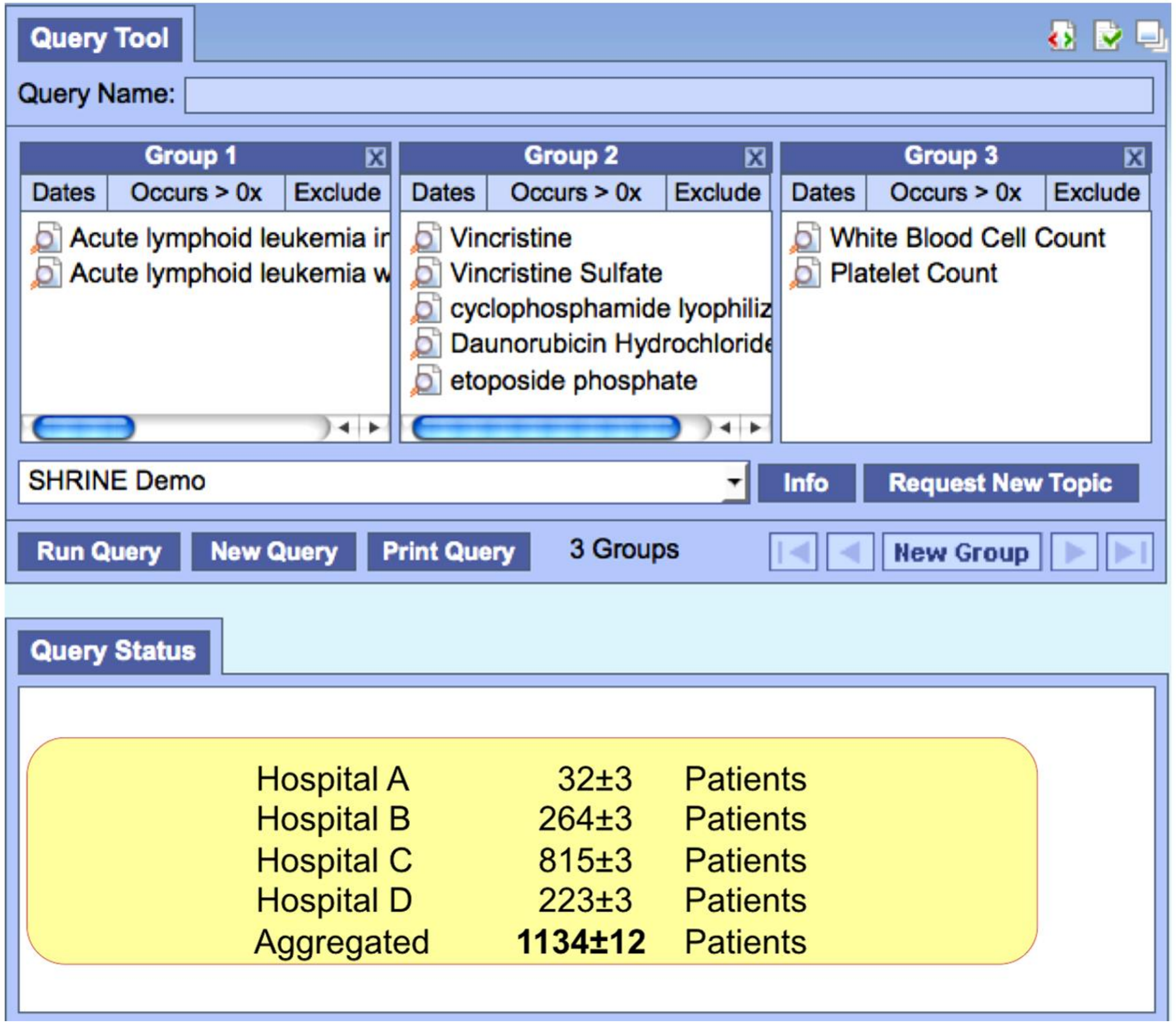
- Virtual Data Warehouse (VDW) maps disparate clinical data systems of HMOs to a common data model (Hornbrook, 2005; Ross, 2014)
  - Acknowledged non-interoperability of inter-institutional data
  - Human layer added value with security, IRB, etc.


Individual Health Care Systems
Administrative and claims data
GHC
KPNC
Other
KPNW
Advance Work
Enormous gains in efficiencies and data quality are made through investments in advance work.
Research centers convert relevant local data to VDW format
GHC
KPNC
KPNW
Other
Research Projects
Lessons learned

Accessible Research Commons for Health (ARCH)

- Arising from the Scalable Collaborative Infrastructure for a Learning Health System (SCILHS), a PCORNet project bringing together previous projects
  - i2b2 – scalable informatics framework to bridge clinical research data with basic science data to better understand genetic bases of complex diseases (Murphy, 2012; Kohane, 2012; Klann, 2016)
  - SHRINE – federated query system across i2b2 data sources (McMurry, 2013)
  - SMART Platforms – substitutable apps, combined with FHIR (Mandel, 2016; Mandl, 2019)
  - Uses completeness tracking system (CTX) to assess data completeness (Estiri, 2019)

SHRINE i2b2 query tool



Group 1:
- Acute lymphoid leukemia in
- Acute lymphoid leukemia w

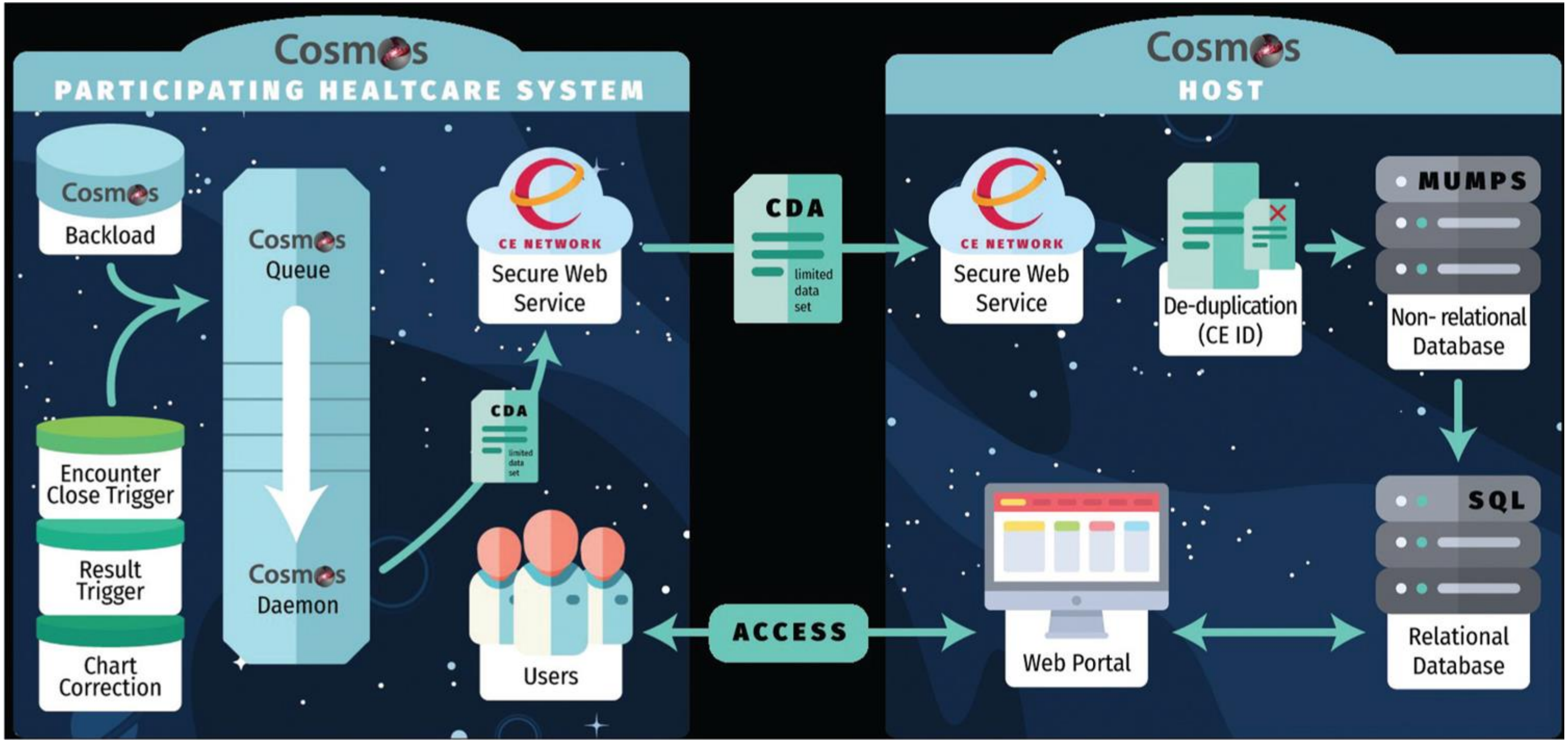
Group 2:
- Vincristine
- Vincristine Sulfate
- cyclophosphamide lyophiliz
- Daunorubicin Hydrochloride
- etoposide phosphate

Group 3:
- White Blood Cell Count
- Platelet Count

SHRINE Demo

Hospital A	32±3	Patients
Hospital B	264±3	Patients
Hospital C	815±3	Patients
Hospital D	223±3	Patients
Aggregated	**1134±12**	Patients

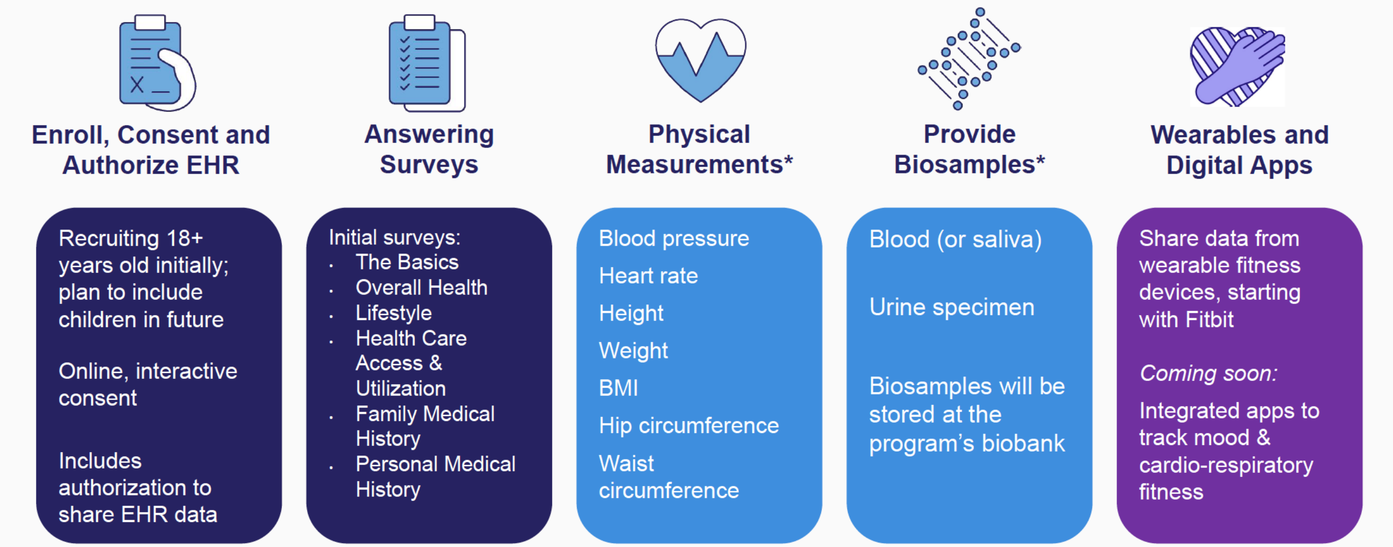
Cosmos Collaborative (Tarabichi, 2021)



Cosmos Collaborative data variables (Tarabichi, 2021)

Concept	Discrete Data Variables
Demographics	Legal sex; gender identity; birth date; race; ethnicity; zip, county and state of patient; date of death; status of patient (alive or deceased); cause of death; gestational age at birth; language (spoken and preferred)
Encounter details	Start/end date and time; type, specialty; reason for visit; age at encounter; pregnancy status at encounter; place of service (zip, county and state); mode of arrival; discharge disposition; organization type
Problems	Diagnosis, including date noted and resolved
Diagnoses	Encounter based admission and discharge diagnoses; surgical diagnoses; visit (encounter) diagnoses; billing diagnoses
Surgical history	Procedure, date/time
Social history	Smoking status, duration and intensity; smoking start/stop dates; sexual activity, alcohol usage status; illegal drug usage status
Family history	Problem or pertinent negative; relationship to patient, age of onset, sex and status (living or deceased)
Outpatient medications	Medication name, type, dose, unit, route, frequency, dispense quantity, refills, and start/end date; indications of use
Allergies	Date noted; allergen; reaction; reaction severity; last updated instance
Immunizations	Immunization; administration date; route, dose; unit
Vital signs	Date/time; blood pressure; pulse; temperature; respiratory rate; oxygen saturation; height; weight; body mass index; head circumference.
Results	Procedure; date/time; specimen source; value and units; abnormal flag; reference range Microbiology organism, sensitivity and testing method if applicable
Procedure	Start/end date; procedure instant; billed procedure; provider specialty
Inpatient medications	Medication name, type, dose, unit, route, and start/end date
Birth data	APGAR score at 1, 5, and 10 min; nourishment method; delivery method; hospital days; birth count and order (if multiple)
Social determinants of health	Social connections; physical activity, stress; education; food insecurity, financial resource strain; intimate partner violence
Insurance	Medicaid, Medicare, privately insured or self-insured status

AllOfUs Researcher Workbench (Ramirez, 2022)



Enroll, Consent and Authorize EHR

Recruiting 18+ years old initially; plan to include children in future

Online, interactive consent

Includes authorization to share EHR data

Answering Surveys

Initial surveys:

- The Basics
- Overall Health
- Lifestyle
- Health Care Access & Utilization
- Family Medical History
- Personal Medical History

Physical Measurements*

Blood pressure

Heart rate

Height

Weight

BMI

Hip circumference

Waist circumference

Provide Biosamples*

Blood (or saliva)

Urine specimen

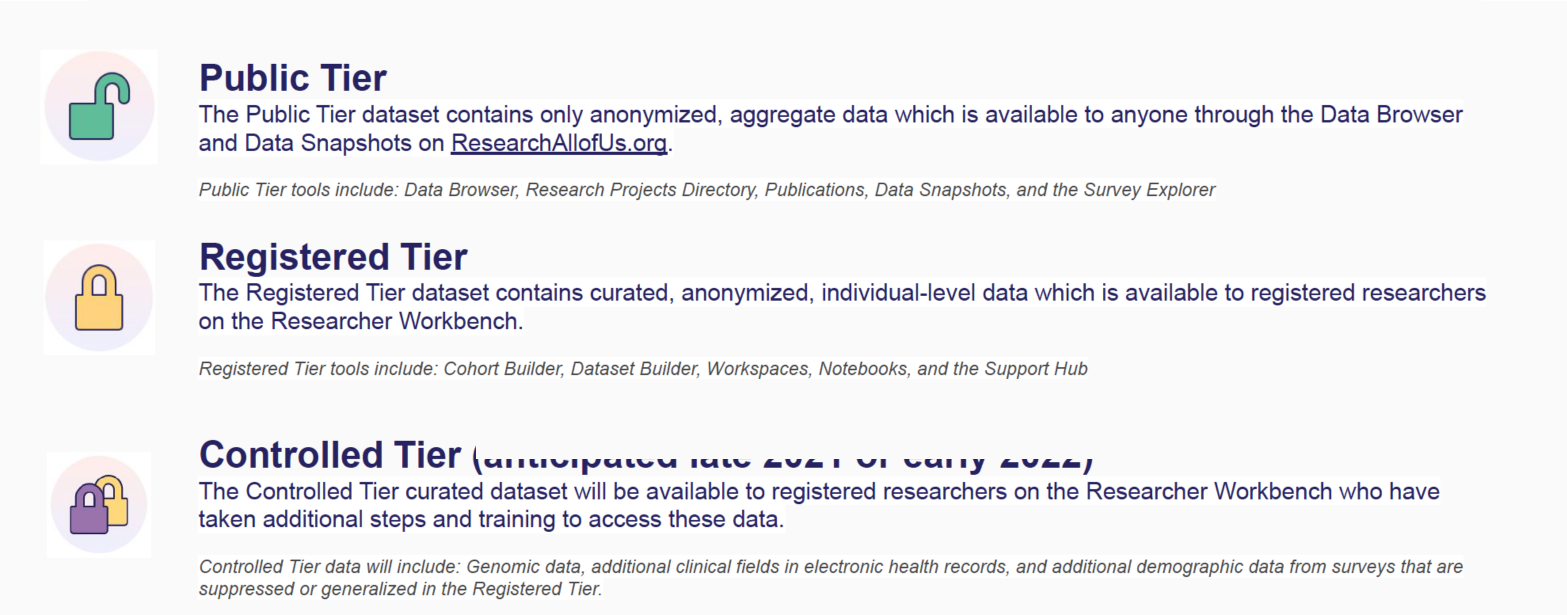
Biosamples will be stored at the program's biobank

Wearables and Digital Apps

Share data from wearable fitness devices, starting with Fitbit

Coming soon:

Integrated apps to track mood & cardio-respiratory fitness



Public Tier

The Public Tier dataset contains only anonymized, aggregate data which is available to anyone through the Data Browser and Data Snapshots on ResearchAllofUs.org.

Public Tier tools include: Data Browser, Research Projects Directory, Publications, Data Snapshots, and the Survey Explorer

Registered Tier

The Registered Tier dataset contains curated, anonymized, individual-level data which is available to registered researchers on the Researcher Workbench.

Registered Tier tools include: Cohort Builder, Dataset Builder, Workspaces, Notebooks, and the Support Hub

Controlled Tier (anticipated late 2021 or early 2022)

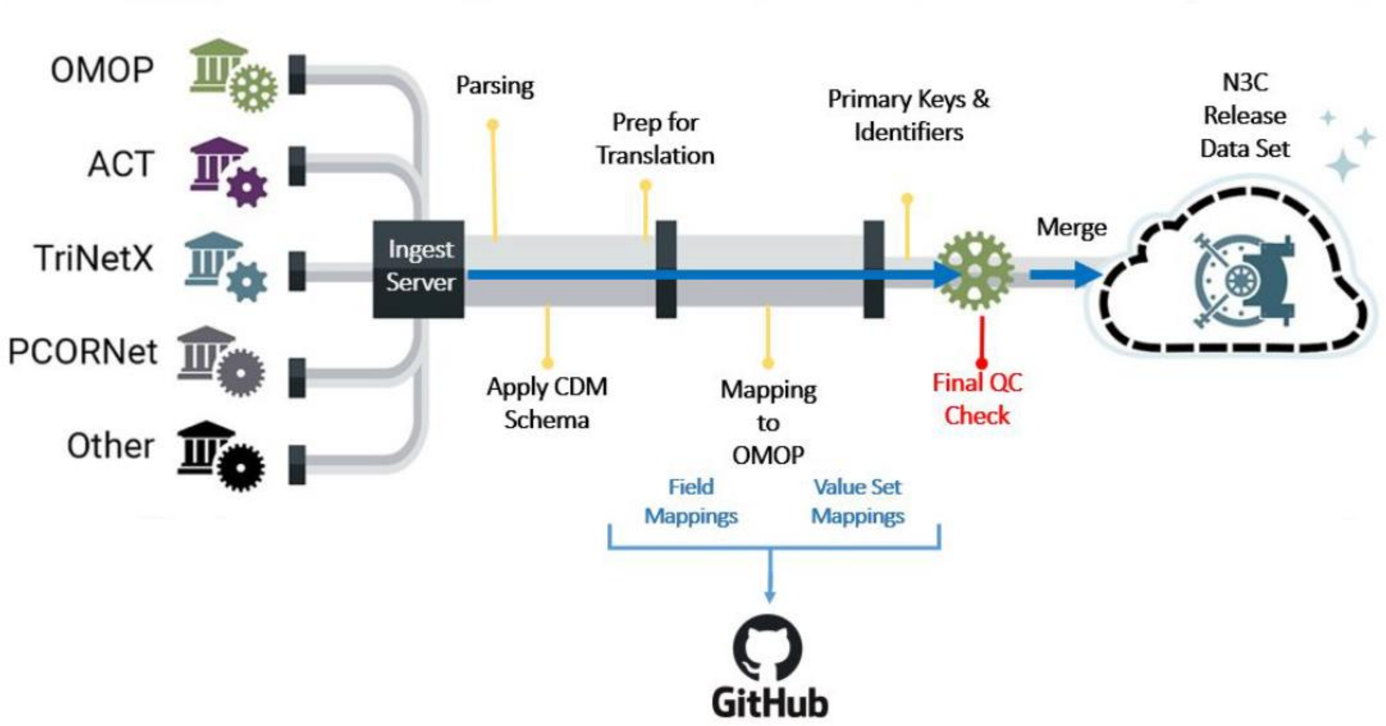
The Controlled Tier curated dataset will be available to registered researchers on the Researcher Workbench who have taken additional steps and training to access these data.

Controlled Tier data will include: Genomic data, additional clinical fields in electronic health records, and additional demographic data from surveys that are suppressed or generalized in the Registered Tier.

National COVID Cohort Collaborative (N3C)

- Outgrowth of CD2H, launched early in pandemic (Haendel, 2021)
- EHR data in OMOP format
- Phenotype includes patients with
  - Positive COVID test (PCR or antibody) OR
  - ICD-10-CM code of U07.1 OR
  - Two or more COVID-like diagnosis codes (ARDS, pneumonia, etc.) during the same encounter, but only on or prior to 5/1/2020
- Each patients demographically matched to two patients with negative or equivocal COVID tests

N3C data overview



Row level data:

- Person
- Drug
- Procedure
- Condition (diagnoses)
- Measurement (labs)

Types of encounters:

- Inpatient, ED & Outpatient
- Longitudinal back to 2018

Bringing IDRs together

- Accrual to Clinical Trials (ACT) – standardizing i2b2 interface to IDRs at CTSA sites (Visweswaran, 2018)
- TriNetX – connect research sites with drug development efforts across world (Topaloglu, 2018)
- Mapping across data models and formats
  - OMOP and OHDSI to FHIR (Jiang, 2017)
  - i2b2 to PCORnet and OMOP (Klann, 2018)
  - Transforming i2b2 data into OMOP in AllOfUs (Klann, 2019)
  - PCORnet to OMOP (Yu, 2022)
  - Clinical Document Architecture (CDA) to OMOP (Katsch, 2025)
  - Use of OMOP CDM for cancer research (Wang, 2025)
- Use of FHIR for clinical research (Duda, 2022)

Examples of research results from IDRs – UK Biobank

- Greater adherence to lifestyle-based recommendations associated with reduced risk of all cancers combined and of breast, colorectal, kidney, esophageal, ovarian, liver, and gallbladder cancers in cohort of 95K people (Malcomson, 2023)
- Sustained dietary change from unhealthy dietary patterns associated with 8.9 and 8.6 years gain in life expectancy for 40-year-old males and females, respectively (Fadnes, 2023)
- Risk of venous thromboembolism in oral contraceptive users and role of genetic factors from prospective cohort study of 240K women (Lo Faro, 2024)
- Improved prediction of coronary heart disease using genetic, social, and lifestyle-psychological factors (Naderian, 2025)

Examples of research results from IDRs – other

- In large academic center in New York City, use of four frequently prescribed drug classes – antihypertensive drugs, statins, selective serotonin reuptake inhibitors, and proton-pump inhibitors – associated with reduced progression from mild cognitive impairment to dementia (Xu, 2023)
- In University of California Health System, many observations for second-line pharmaceutical treatments used for patients with Type 2 Diabetes (Vashisht, 2023)
- Multi-cancer risk stratification from Danish registries (Jung, 2024)
- Discovery of associations between prescribed drugs and dementia risk (Underwood, 2025)
- Cardiovascular post-acute sequelae of SARSCoV-2 in children and adolescents from US EHR data (Zhang, 2025)

Other informatics innovations to support clinical research

- Research Electronic Data Capture (REDCap) uses metadata to create simple data capture forms and stores data securely for researchers (Harris, 2019)
  - Recent integration with FHIR (Cheng, 2021)
- Clinical Data Interchange Standards Consortium (CDISC) – standards to support electronic acquisition, exchange, submission, and archiving of clinical and non-clinical study data and metadata
- Increasing role for generative AI, e.g., mapping local data elements to NIH Common Data Elements (CDE) (Wang, 2025)

Some lessons learned

- AllOfUs
  - Modest agreement between survey data and EHR – low for conditions not commonly reported in EHR, variable for different disease categories – highest for cancer (Sulieman, 2022)
  - Among 2515 diseases and phenotypes in both AllofUs and UK Biobank, 86% had higher prevalence in AllofUs than US general population or UK Biobank (Zeng, 2024)
- N3C
  - Able to harmonize many but not all units and values across sites (Bardwell, 2022)
  - Able to associate demographics and co-morbidities with higher clinical severity and develop machine-learning models predicting disease severity (Bennett, 2021)