Via the break out notes from the Federated Data Query Workshop, here are the topics that were discussed and they fall into three or four suggestions.  Chris, I think you are focused on the Outline/Manuscript but wanted to include the other items in case they might be weaved in.

Here is the folder of workshop materials as reference: https://drive.google.com/drive/folders/1kgcoV8BW_9Zg7XLwObw9HXAO4aJLamQQ?usp=sharing

Also change the title as you see fit.

- Outline/Manuscript
- "A Short History of Common Data Models" using Chris's presentation as a starting point
- Technical report/guidelines
- google doc for this meeting enumerating the tasks, functions and deliverables important to them

# Clinical Data Harmonization and Federated Query for Translational Research:

# Reflections and Report on a CD2H Workshop

Authors: [if you believe you made a substantial contribution, add your name.  We will make them alphabetical save first and last]

Christopher G Chute, Johns Hopkins University
Jonathan C. Silverstein, University of Pittsburgh
Guoqian Jiang, Mayo Clinic
Shyam Visweswaran, University of Pittsburgh
Emily Pfaff, UNC Chapel Hill

The outline below is tentative. Happy to entertain suggestions for additions/modifications/re-ordering. The default permission on this doc is Suggesting. Let me (Chris Chute; chute@jhu.edu) know if you are unhappy with that.

# Abstract

# Introduction: Vision of CTSA Hubs as an Integrated Network for Translational Science

# The Case for Comparable and Consistent Clinical Data

# An Overview of Clinical Research Data Models

## OMOP/OHDSI

## PCORNet CDM

## ACT

The Accrual to Clinical Trials (ACT) network is a federated network that consists of local Integrating Biology at the Bedside (i2b2) EHR data repositories that are linked by the Shared Health Research Information Network (SHRINE) federated query tool. The network uses a CDM that specifies data domains and data elements to be loaded at each site's i2b2 repository and a common query ontology that is installed at each site's i2b2. The ontology enables a user at any site to construct queries that are executed in real-time across the network. ACT supports a variety of research needs including assessing study feasibility in terms of patient counts and cohort discovery.

## I2b2 Star Schema

The i2b2 repository employs a star schema that consists of a fact table that stores biomedical observations and is linked to a set of dimension tables that enable grouping of observations (e.g., by patient or visit). The star schema is sufficiently flexible to enable the implementation of commonly deployed Common Data Models such as the PCORnet CDM and the OMOP data model.

The i2b2 Star Schema was developed as a CDM that enables conformant transformation of patient data to a common data structure and representation of meaning. i2b2-based solutions have been widely used in clinical research communities such as the Shared Health Research Information Networks (SHRINE) and PCORnet. Building on the i2b2 framework, the tranSMART platform is an analytic platform that also incorporates the ability to load molecular datatypes, including those derived from next generation sequencing (NGS).

## CDISC BRIDG

In 2004, the Clinical Data Interchange Standards Consortium (CDISC) initiated the Biomedical Research Integrated Domain Group (BRIDG) project in collaboration with FDA, HL7 and National Cancer Institute (NCI). The collaboration’s effort developed a domain analysis model that is a shared view of the dynamic and static semantics for the domain of protocol-driven research and its associated regulatory artifacts. The goal of the BRIDG model is to provide an overarching model that could readily be understood by clinical research domain experts and that would provide a semantic basis for harmonization among standards within the clinical research domain and between biomedical/clinical research and healthcare.

The BRIDG model has been used in a number of research applications as a foundation to enable semantic interoperability between clinical research and clinical care domains and beyond. These research applications include 1) the harmonization with the OHDSI CDM to enable linking the protocol driven clinical research artifacts to active post market surveillance studies, 2) the development of the Life Sciences Domain Analysis Model (LS DAM) to support translational research through shared semantics across the life sciences and clinical research domains, and 3) the EU-funded projects including SALUS and TRANSFoRm to create a learning healthcare system at a European level to support multiple types of research using primary care data, to recruit and follow patients in clinical studies, and to improve diagnosis and treatment.

## FDA Sentinel CDM

The Sentinel Operations Center (SOC) coordinates the network of Sentinel Data Partners and leads development of the Sentinel Common Data Model (SCDM), a standard data structure that allows Data Partners to quickly execute distributed programs against local data. The SCDM was developed in accordance with the Mini-Sentinel Common Data Model Guiding Principles and was modeled after the HMO Research Network Virtual Data Warehouse. The SCDM currently includes 12 tables that represent information for the data elements needed for Sentinel activities.

TriNetX

# Clinical HIT Standards

## HL7 V2

## HL7 FHIR

HL7 Fast Healthcare Interoperability Resources (FHIR) combines aspects of HL7 v2, v3, and CDA into a new standard with a focus on ease of implementation. FHIR’s building blocks are “resources,” like Patient or Encounter. Individual users can adapt resources to their needs by combining them into “profiles,” which can be further customized with “extensions” to make use of data not covered in the standard FHIR resources.

## US Core Terminologies

## openEHR

openEHR is a multi-level data model. At the bottom, a stable "reference model" (RM) contains highly generic "building blocks," or data structures that can be used by higher-level classes (e.g., the concept of a DATA_VALUE). More specific domain concepts are expressed in higher levels, as "archetypes." An archetype is a collection of RM building blocks given specific meanings to model a given domain concept. (E.g., a blood pressure is a concept made up of two numeric data values with the meanings "systolic blood pressure" and "diastolic blood pressure," respectively.) Adding a new archetype simply requires combining and defining the right RM building blocks (and does not require any changes at the RM level or its persistence mechanism), which makes openEHR highly adaptable to changing data needs.

# Federated Data Query

## Aggregation vs. Decentralized

## Biosurveillance (CDC-public health)

## Post-marketing (FDA Sentinel)

## South Carolina network

## SHRINE and ACT

The ACT network is federated where each site manages and controls its i2b2 repository, and the repositories are linked with the SHRINE federated query tool. SHRINE allows querying the network in real-time; this allows a user to iteratively refine a query based on results obtained from previous queries. ACT uses a common ontology across the network that allows a query to be constructed using a common set of concepts. Queries in ACT represent definitions of computable phenotypes that are both human-readable and computer-executable in any i2b2 that has the ACT ontology installed.

## PCORNet

## TriNetX

## Feasibility of a Common Software Stack

# FHIR as Canonical Hub

## Role of FHIR Repository

## Leveraging of Research CDMs and Communities

## Overview of some FHIR Servers

### HAPI/CDR Smile

### Microsoft Open-source - Supported

### Google FHIR server

### Cerner Bunsen

# Action Items from the Workshop

## Establish Connection between HL7 and the CTSA Communities

## Conduct Systematic Study of Gaps between FHIR and Traditional Research CDMs

When mapping between a traditional research CDM and FHIR, there is potential for loss of data (where source data has no good equivalent in FHIR), change of data meaning (where FHIR equivalents are close, but not an exact match), or loss of granularity (where FHIR value sets have less detail than source value sets). These issues are not uncommon in data transformation in general, and are certainly not limited to transformations to FHIR. It is important, then, to put data mapped to FHIR (or any transformed data) in its proper context. For a given use case, if highly granular detail from the source system is important to the research question and that detail is lost during transformation to FHIR, then FHIR may not be sufficient in and of itself for that study. In short, no data model is the right choice for all applications, but our hope is that FHIR's breadth of data domains and wide adoption would allow it to serve a large variety of use cases.

## Identify and Package FHIR Education Materials for CTSA Audience

## Conduct Deeper Comparisons among FHIR Servers for CTSA Suitability

## Convene Task Group for Informatics Sustainability and Change Management in CTSAs

Mapping EHR data to any one of the major CDMs is resource- and personnel-intensive, and requires an ongoing commitment to maintain and refresh infrastructure and data over time. As institutions are asked to adopt more CDMs (and the number of available CDMs multiplies), this level of effort increases and can quickly become untenable, even with existing expertise and education/documentation.

# Discussion

Given the substantial effort and substantial progress we anticipate in the clinical arena standardizing on FHIR, the research informatics community is exploring leveraging FHIR in clinical research. While this is ongoing we must consider the different needs of research relative to clinical, particularly in regard to bulk data usage, which is limited in focus within FHIR. Similarly, the need to consider harmonizing approaches to ETL from EMRs for efficiency and consistency among academic health systems is a strong opportunity area. If we ignore the local variations in approaches to ETL, data management and data labelling and focus only on consistent formats and structures, not consistent data, we will have interoperable data, but not the right scientific conclusions. As such we may want to consider a wider set of action items with the focus being upon reducing variation across the enterprise of enterprises that make up the research community. Thus, in addition to trying new things, sharing best practices may be an important pathway forward.

# Conclusions

So as not to disrupt the flow of the document, but also to surface my substantial notes I place them following here assuming someone will use and/or move them:

CD2H - Federated Data Query Workshop

Karthik Natajaran - OHDSI
Hundreds of millions of records worldwide
Many international meetings, governance, voting processes
Many are de-identified without dates
OHDSI Coordinating Center does collection and aggregation of queries and analysis - federated query model with human assent at each site
Many tools now - ATLAS - really a researcher warehouse - cohorts/data are SQL built and open source sharing of cohort sharing across about six flavors of SQL
Open source community that allows sharing of all components - code sharing (R packages, etc..)
Creating evidence resources such as howoften.orrg - a work in progress - designed to clarify frequency of side effects for active ingredients
All of Us - Big Query in Google Cloud
There are efforts in OHDSI to map FHIR into OMOP model - with this, a FHIR feed could create OMOP data and OMOP could return FHIR - this has potential, but figuring out what that Federated query looks like would be helpful, particularly in batch style

Harold Lehmann - PCORnet
Clinical research
PCRF, NIH, Pharma
Junior faculty, PIs, informatics, IT, clinical, researchers
11 networks
pcornetcommons.org
210 queries across 75,000,000 patients
Tends to be top down
Pragmatic and Observational studies - list shown of PaTH-involved studies
Model of federated query
Summary queries can have tens of thousands of result rows due to stratification
Daquery - row level query
200+ data characterizations and quality checks
CDM 4.2
Description of data model and processes for converging
Discussion of Codes
Michael Kahn’s harmonization framework with feedback in red and blue responses
Each site responsible for data cleaning wordlist - about a half FTI per site
PCORNET_TRIAL table
PATH_TRIAL table
Analytic Issues

> “Prior experience not needed to use CDM” - not true

Question - can studies be done from CDM or is more detail from the EMR required in every study

Philip Trevvett - ACT
Description of i2b2, local as well as Shrine
Data Domains included in ACT Data Model
ACT Ontology Goals - growth of domains, currentness, historic completeness, forward compatibility, usability, derive ontology from UMLS - additional work to include deprecated codes including discoverability
Data Coverage - Data Characterization Overview (e.g. its a Children's Hospital, or certain sites don't have ICD10 Procedures)
Weekly Smoke Test (example shown and can see missing or inconsistent data to follow up upon, also date limits tested), and Characterization Survey before joining the network (more detail presented on this also) - what present and absent, what areas focused on, data growth, etc..)
Heading toward Data Quality Assessment
Understanding Query Results - each woman has peculiarities that a user needs to be aware - sometimes deeper ontology understanding is required - some issues where user doesn'tt have enough context - example shown of query for presence versus presence for value - can show no counts where there are not values loaded (different result for presence versus value)
Another example - Data Mapping Issues (specific to medications) - mapped to ingredient versus more granular SCD one gets zero results - local data control issue
These examples show its not just data quality and completeness among sites but how to convey to users what's in there and how to use it effectively
Medical Devices asked as question - so far we're putting in CPT and it becomes a site decision to a degree
ICD10/ICD9 convergence - current state and working toward autoupdating ontology from UMLS

Ken Gersing - CDM Harmonization (CDMH) …many beyond just those three…FDA, NLM, ONC, NCI, NCATS, CDC, CD2H
PCORTF Project - driven by contract with FDA suggested by FDA commissioner Rob Califf - motivation was line-level data, not aggregate numbers
Phase 1 - harmonization - radioshack CDM "adapter model" concept shown for researcher writing one query and it going to the four major models (above plus Sentinel) - shows diagram of adapter model in some detail with query transformation and causing ETL "save as" csv or FHIR
Many things built - mappings to FHIR and BRIDG (CDISC to go to FDA) and more in first 2 years
Phase 2 - planned to use FHIR but it wasn't ready in the community so many transformations to many versions to many databases (the "language translation" didn't work so well because of geometric complexity of models*versions*databases)
Phase 2 - identify areas of work and solutions assuming all the models are good - ***Ken's key realization is that the trust relationships and the networks built are much more important than the technology***
Bring data into a "FHIR server" possibly
Shows interesting Phase II "Clinical Adapter - Data Flow Diagram" - agencies in green with sites in blue with FHIR in between (either pulling into FHIR database routinely or pulling the model directly in and then in "agency" converting to FHIR)
Note that FHIR is clinical standard rather than research standard which means there is much Federal and site budget behind it
Phase 2 is 4 or 5 months in
Phase 3 is to "put CDS hooks" into the FHIR model
Getting other health care standards agencies and Federal agencies and big tech industry and EMR vendors engaged for adoption by all
FHIR is on Fire - adoption is amazing, really happening - we should take advantage of this
CQL Tools (engine, runner, execution) - Mitre/AHRQ - SQL-esque or CDS authoring GUI tool - to generate FHIR query against a FHIR "server" (many open source options, not chosen one yet) - we don't yet know if this is scalable
12-18 months we'll have most of this running in prototype

Question is “have you validated against repositories yet” - yes, in phase 1 we did - we have not done that with CQL yet
Comments that FHIR from EHRs is not robust at this time - bypassing the architectures and going with FHIR from EMR is concerning to audience questioner
Ken response: there are many issues with retrospective data and we don’t yet know yet if we’ll use data back or choose a forward date or how it will really work - so far its a “technology project” and route and form in RxNorm for example (comment that “there is no route in RxNorm” from its creator)

Harold Solbrig - i2FHIRb2 - Lessons learned putting FHIR in i2b2
I2b2 Cell model description with pluggable and replaceable APIs and intercell communication
Shows web client cell - one of many cells
Talks about the ontology in i2b2 - considered to be a representation of an information model for clinicians
Star schema data model - atomic fact table
Ontology cell - assembles SQL on the fly to go against the fact table
I2b2 is VERY flexible - easy for researcher to create model - bad news is that two researchers can come up with two models for same question that are very different (and underlying data and ontologies can be different too - this being worked on in ACT)
ACT ontology growing, but mostly growing slowly just for i2b2
Could we work on FHIR data model and put it into i2b2 ontology in a way researcher could actually use it?
Use the RDF resource representation in FHIR as each separate item as fact - it has URI for each structural core - so there are concept identifiers for everything one can put in FHIR
We mapped FHIR RDF into entries in i2b2 observation facts table (FHIR has notion of resource that i2b2 doesn’t)
We were able to transform patient data from FHIR to i2b2 (FHIR also has non-patient data, providers, etc…that we didn’t map) but would the FHIR ontology model make sense to researchers - answer: NO - the FHIR data model becomes exposed
***Is there an inexpensive way to correct this? Possibly yes: make it conform to the ACT ontology! - we decoupled the FHIR data model from the information model and data model - then FHIR resource and ACT queries could be blended - novel information available immediately***
Conclude: we can put FHIR patient focused resources and in to the i2b2 data model - it was non-trivial because there are patient-focused aspects in i2b2 - e.g. FHIR metamodel could map to i2b2 “model” (not hard because of i2b2 data model)
What we really need is Common Logical Clinical Information Model - FHIR is how to exchange data not to store it - CDMs are about how to store it - ACT ontology is a common logical model which is more of what we need (really a logical information model needed that can be realized in various different and useful physical models)
Still waiting for 20 years since beginning of HL7 v3 RIM to realize this
Need to talk about storage models, data interchange models, and logical models (the logical models provide the semantics)
Logical models can be used beyond research for clinical too (CDS)
Comment from audience - where is the information that we’ve lost in data?  Problem of deciding “what is atomic”? Perhaps this depends upon our level of discourse? - answer: part of the discussion in the history of this was “what is an atomic piece”, but we need to use context too - we should possibly talk about common information “models” not one

Chris Chute - FHIR as a canonical data model
Starts with LHS vision of Chuck Friedman
CTSA hubs are a network leading and funding this
OHDSI is also quite big and enormous contribution
PCORnet also quite substantial
TriNetX
ACT network - arguably CTSAs answer to Federated Query

FDA Sentinel Network started 2008
CDC PCOR Trust Fund
What do distributed query networks actually require?

- CDM
- Semantic Binding
- Authentication Layer
- Authorization configuration
- Query Broadcast and aggregation

Perhaps we need a “shared stack”? Is that possibly as an “appliance” if you will that will serve many networks?
Currently we maintain many models (all of those above at Hopkins) including next-gen FHIR repository) - tiring managing all these
Is it possible to have a repository that can do these all? What’s wrong with using the clinical data models? (Answer in research community historically is clinical data model didn’t really exist so many many research ones created and evolved)
LHS circle model with data coming from clinical data - knowledge - back - standards, comparability and consistency holds it together in the center
Has clinical community now caught up with a model - with FHIR? FHIR has emerged over past five years…FHIR quality is not yet great from EMRS…FHIR endorsed across all stakeholders (government, industry, payers, academic, community) - clinical support in terms of finance dwarfs any research efforts - FHIR resources being drafted even for genomics - why wouldn’t research community leverage all that as primary
Think of FHIR as logo pieces with micro schema for data marts - FHIR microschema (demographics, observations, medications, procedures) using graph model - discrete data elements with just-in-time model binding
Data Element Soup idea - can include traditional
Is FHIR ready to be canonical Hub CDM? Not yet (still in draft), a messaging mechanism not a model (has this de facto made a model), deliberately underspecified (semantically)
Pros: bindings can be done (nationally mostly) - can we have a shared semantic binding for research? EMR can’t function as native FHIR (for auditing, transformation, performance) so we do need a research set of resources - the physical model behind it is irrelevant actually (SQL, JSON) as long as its conformant and performant
Who embraces this: a lot of folks - but complex query (joins and other cross-population) being worked on - not yet FHIR Bulk, FHIR Bundle, FHIR SQL, or FHIR CQL but being worked on
What about traditional CDMs - they will not go away - they could be facilitated interoperability with FHIR “Hub” - big diagram with FHIR in the center with pointing in from using agencies, data models, clinical and research
FHIR is a compelling candidate

Leslie Lenert - Federated on FHIR Data System
MUSC and Health Sciences South Carolina - 3 research universities and 6 health systems - statewide-ish CTSA
Focus today on multi-institution EDW approach - having to maintain all the data models drives this presentation - all those models plus HIE and Quality too
Architecture of EDW - health systems collecting - consolidated mapping and storage (including statewide EMPI) - analysis layers with marts and registries
This architecture is expensive and difficult to maintain - not useful to our clinical members which was a plan but the clinical systems build their own data warehouses so we became not so important - more used higher quality it gets so this was somewhat problems
Showed model of functional model of operations (big swim lane flow diagram of all the ETL) to the board who concluded this is complicated (duh) and not a board level - part of the point is that there is a lot of complexity people choose to be ignorant of I think was his point - so one can have a single architecture, but leadership may not care or directly endorse
HAPI is open source FHIR server front end interface underlying most FHIR development
The authors of much of HAPI protocol built an infrastructure called SMILE-CDR
Thus, there is an evolving commercial product that does do this centralized data hub model SMILE-CDR

So Federated on FHIR is a network of FHIR CDRs - link individual site FHIR repositories to central FHIR repository - each linked outward that does dynamic transformation to the desired format for output (FHIR is the lingua franca but we're doing automatic transform to multiple data models)
We use the FHIR Subscription Resource - this is a push-based resource that sends when field is triggered out to other models (e.g. PCORnet, OMOP, i2b2)
Subscriptions to this service includes an EMPI for integrated record
The Subscription feature was really designated for clinical but can send emails for eligibility for trials (e.g. so-and-so ADT says in hospital, notify researchers)
Clemson data center has Cloud-based Implementation of this - SaaS Multi tenant model
Loaded 2-years back-loading - some flat file, some ADT archive messages loaded
Summary: Federated on FHIR is robust cloud-based platform for translational research - can directly transform health system data into canonical model - creating CDM and ACT, but have created OMOP already - do merge across institutions

John Loonsk - Commonalities between public health surveillance and federated query approaches - CDC folks - CMIO for public health labs - consultant to CDC - also present Arun Srinivasan
Public Health Surveillance related to what "you" do
Public Health much smaller than research with is much smaller than clinical
We need to use the burgeoning clinical model - bought in to FHIR early on
Centers CDC is made of
Identification and monitoring of health conditions and events as well as their investigation and management - work with identified in some circumstances and de-identified in the circumstances - we're "all site" including community providers
High level requirements

- full automation of reporting
- reduce burden on providers
- leverage infrastructure of others

We think there are common needs to develop surveillance AND distributed query needs
We value need for de-identification, and authorized re-linking, supplemental data acquisition for completeness and in investigation - we see these as services
Clinical data models have been inadequate - but we have moved early in newer ones - according to federal regulation we are to use new U.S. CDI approaches, but also interested in greater specificity and quality we need
We see opportunity to collaborate (my summary of a fair amount of more presented thoughts by John)

Mark Ciriello - Harvard Catalyst - ACT Network and SHRINE Development - Bill Simons for technical backup
ACT Network concept and technology - example of ACT query execution

- background of ACT
- nice diagrams presented of the network model (what is SHRINE, what is at sites, etc…)

ACT Network operations - lessons learned

- transparency of availability - and of course desire high availability
- network composition is changing (more sites all the time)
- need credible results (there are non-obvious end user understanding required for accuracy and completeness)
- needed clear guidelines of system administrators at sites
- needed network operations work group - complement to other ACT work groups - formalizing procedures and operational standards

Network choices made over setup process

- Production network, Stage Network (local testing and new nodes joining), Test Network (leading edge)
- weekly maintenance window
- central network operations support

help desk ticketing system (tech, shrine, ontology, project management, i2b2 data, etc…)
Network site connectivity status (network monitoring) - currently smoke tests, next rev will be optional configuration of network health dashboard showing connectivity (related to data characterization and dissemination to end users)
Audience question on how we went about data characterization - Phil Trevett answers regarding Pitt doing survey phase 1 of what at high level at each site and its general content plus working toward phase 2 in process

Casey Overby Taylor - eMERGE EHRI WG - focus on genomics for today (also Mullai Murugan presenting)
Sequencing centers genomic rendering with FHIR - return of results to patients focus in this phase
PDF results from labs do not facilitate CDS, etc…
All sites sending to two common sequencing centers (Baylor and Partners)

- raw data return
- focus on clinical reports returned in structured format

Many sites returning results but local infrastructure needed with ancillary genomics systems (e.g. integration with Epic and other projects - e.g. at Northwestern)
What is needed

- Computable clinical sequencing results (sequencing studies)
  - not yet established national standards for the **data**
    - in survey found some sites were parsing PDFs so we needed to standardize - FHIR was not ready then - went with industry GeneInsight “standard” - its worked well
    - in the last year working toward national standard - we don’t want to stick with proprietary to eMERGE - looked at normative release of FHIR Core/CG but not enough support (just sequence) - there is clinical genomics workgroup for FHIR but still in draft mode - decision to go with clinical CG FHIR workgroup even though very early - reconciling with the CG group now - creating FHIR spec for eMERGE, pilot implementation, and anticipate reconciliation with CG FHIR spec
    - shows big mapping process diagrams to FHIR
    - challenges noted but not enough time to present them today
  - local **delivery** mechanisms too not standard
- Infrastructure (health IT)
- Engagement (study team)
- innovation opportunities
- Governance and sustainability

Hopefully getting to national standard

Mark Overhage - Cerner: Really Big FHIR
Cool that this is meeting of geeks
Lots of ways Cerner is leveraging FHIR
Will talk about the Big Data research parts of Cerner
HealtheIntent Platform - a centrally managed but segregated (with aggregation and normalization, apply an intelligence layer, measure it, and push to clinical workflows)

- 1000 data connections
- 60 connected EHRs
- 18 PB of data

HealtheDataLab - usual data science tech (AWS, Jupyter, Python, R) attached to HealtheIntent Platform - key is data model understandable documented and available to researchers - we have been happy with FHIR data model as standard because of community out there to explain data model so we don’t have to - FHIR has the structure for deeper data (like who the technician is in the lab, or is this a verified diagnosis - so its in there, but not always populated) - killer app is that with FHIR you as data scientist can push button to “deploy” in production - we’ve automated that because the data models match with FHIR (10 minutes from validated model to production)

Open source Bunsen project - on GitHub - FHIR first class integration into big data model with Apache Spark - using FHIR resource definition - at run-time you are exposing efficient set of data structures from FHIR to Spark (not sure I understood here - went very fast in Jupyter notebook-like layout with Python) - also includes distribution across nodes, so not limited to one big server

This also is married to CQL as an “expression for portable knowledge” - electronic clinical quality models (eCQMs) and CDS - Hedis, etc…

Internal and External validation of models then push back into computational engine

Chronic disease models with multiple machine learning tools - run in 15 minutes - send training across thousands of nodes in AWS and get results

Vivian Neilley - Google - Ingesting and Harmonizaing Clinical Data Using Google Cloud

Google Cloud is about giving tools to individuals

Many examples I’ll show is how we can help you do things efficiently (we are similar to other public clouds)

Ingestion

- Healthcare API (HL7v2, FHIR, DICOM)
  - Also allows for persistence of storage not just ingestion - an OLTP database based in FHIR
  - Integrated with Apigee for multiple APIs management
  - Cloud Storage and BigQuery and GraphQL
  - APIs also allow for de-identification configurations that runs across FHIR store

Harmonization

- Cloud Data Fusion (going GA in September - not yet for live patient data)
  - managed, enterprise data integration service
  - write pipelines without using code
  - graphical environment with hundreds of plugins and hundreds of transforms (e.g. drag and drop ingesting of HL7 pipelines), data quality checks, joins across databases, etc…
  - metadata integration of tags and lineage support of what happened to that specific feature, how it was transformed, etc…
- We’re building many data harmonizations of clinical sources to map to various schemas - working with CDISC, OHDSI for mapping transforms from source to destination schemas

Analytics

- Big Query scales to Petabytes but still allows SQL columnar format
- We support SQL on FHIR
- Many tools Google does to make machine learning possible in Google Cloud to anyone
- We’re approaching this to allow integration back to clinical via API so model can be used clinically
- Pipelines API for genomics, DICOM, and create ML models across modalities

Working to support 100+ data sets

- NIH
- World Bank
- RxNorm
- Human Variant Annotation Datasets
- HCPCS
- Hospitals registered with Medicare
- CMS data (which I missed…)

Nansu Zong - Mayo Clinic - NLP2FHIR: A scalable FHIR-based Clinical Data Normalization Pipeline and its Research Applications

Background of the problem of AI solutions transmitting to other institutions and making them work

Need to improve portability of data

EHR Data structure section detection (SecTag section dictionary) and normalize the section name with LINC codes (FHIR compositional resources) - NLP pipelines (Apache UIMA) - FHIR clinical resources

Many applications

- PheWAS Assessment
- EHR Driven Deep Phenotyping - a classification problem - using deep learning
- FHIR based data presentation
- Graphic tools to show the contributions of different features to deep learning model
- Automatic data capture for clinical trials - nice examples shown - with validation

From BDKonFHIR at Mayo GitHub apparently available for this

LUNCH

Panels: some fairly interesting rapid discussion throughout participants (participants list and all powerpoint and video to be shared soon)
Some insights I gained from them:
Use cases and funding probably more important than technical approach in where this will all go…
FHIR seems to have enabled big tech to engage in clinical informatics (an API for medicine).
Knowledge engineering (developing the phenotypes, features, etc..of relevances) is the biggest total task…
Chris Chute asks: “Is phenotypic THE use case?” But, Gersin says what about genomics, and Silverstein says, what about causality…etc…much discussion
Multiple models and CDMs are fine as long as they are useful (Lennert and Google folks agree)
Canonical model (zero level - atomic) is the way to go? Is that canonical model FHIR?
We MAY as a country follow the LHS model of archipelagos of projects assembling to larger projects (e.g. ACT, PaTH, etc…)
Big science technologies are known to do distributed authentication and authorization, we just need to adopt them where needed (JCS says, Google agrees)