GROUSE Architecture

Introduction:

-Presenter: Dr. Song research interest is in data-centric machine learning models and knowledge discovery algorithms.

-GROUSE stands for GPC Reusable Observable Unified Study Environment.
-GROUSE provides de-identified resources that merges CMS claims and GPC sites EHR data.
-It was created to fulfill the phase 2 requirements of PCORI to in order to assess outcomes for patients who may not remain under a single healthcare, PCORI requires to develop some strategies to integrating insurance claims. The initial strategies of claim integration is to use Medicare and Medicaid claims.

-the 3 overarching goals of creating GROUSE vs characterized the increase of data completeness,
Deidentified resources that merges CMS claims and GPC EPIC data to:

1) Characterize the increase in data completeness and comprehensive provided through claims integration to provide in a more “complete” picture of our patient’s health. Use insurance carriers to find patient migration to get at outcomes
2) Evaluate the distributions of health care processes for patients in 3 focused areas, breast cancer, ALS, obesity (cancer rare, common), understand how studies of the GPC population generalized to the broader populations in our states
3) Using CMS data to support trial recruitment and observational studies. So that we can provide a more complete picture to describe the continues of care for the patient.

For each of the sites, they are required to choose 3 health conditions to focus on, for GPC which host one rare condition which is ALS or Lou Gehrig disease, and a common condition which is obesity, and breast cancer.
-We will use this integration environment to start addressing the distribution of health and care process for patient within these conditions and finding patterns within the GPC for Medicare and Medicaid population.
-The use the CMS claims data as a more representative denominator to enhance quality control process and establishing this correlation with CMS claims data for health for data support, trial requirement and additional studies.

CMS Research identifiable files broken down into 5 parts:
Within the current G environment,
-Medicare Claims:
Part A: for institutional activities (2011-2017)
Part B: for non- institutional activities (2011-2017)
Part C: Managed care, different types of services provided by Medicare (not yet purchased)
Part D: Prescription drug event (2011-2017)
-Medicaid Claims:

MAX/TAF: Medicaid data, has several transformations, not very clean and a lot variation across states. The older data (2011-2012)

Type of Medicare beneficiaries:

1. Elderly
2. Disabled
3. ESRD, ALS

Type of Medicaid

1- Low-income families
2- Pregnant women
3- People of all ages with disabilities
4- People who need long-term care

Strength of CMS data

1- Demographic validity (above 98% of orderly Americans are Medicare)
2- Coverage Size (24 million are Medicare and Medicaid across 9 GPC states)
3- Representative Mortality (has a national death index)
4- Multiple providers
5- Consistence format
6- Availably: Medicare 13mo latency for annual files, can get 4 mo but costs more, Medicaid is about 24 mos latency for annual files
7- exists more than 200 charge and payment related variables (PCORnet not interested in costs, but imagine that this WILL be future questions). Can provide data dictionary for all these columns.
8- 7-financial information
9- They are very detailed

Question:

Alex Stoddard (MCW)
Q: Is the CMS data in GROUSE is nationwide?
A: Not nationwide, but it's covering the 9 GPC states.
The Medicare claims is covering the entire states, and a subset of these data can be linked to EMR data coming from GPC sites.
Additional Q: what to define someone to be in a state, for services of Medicare coverage,
A: Medicare data have those beneficiary basic demographic table, (denotator file) and it uses a column to identify patients coming from a particular state. Possibly the enrollment data.

Jim Campbell (Nebraska)
Q: for example, a patient in Florida as they are insured there, in Nebraska and get care, we provide services and then we file a claim against the Florida carrier.
A: in the denotator file, the patient's residency will be in Florida, but in the carrier claims their NPI number associated with all of the carrier claims or institutional claims so the NPI could be used to identify what are the institutions.

Bradley McDowell (Iowa)
Q: To flip Jim's question, would claims for care that an Iowa resident received in Florida be in the set of claims available for analysis?
A: Yes, as long as they got billed with that particular beneficiary. (**I will follow up with that**)

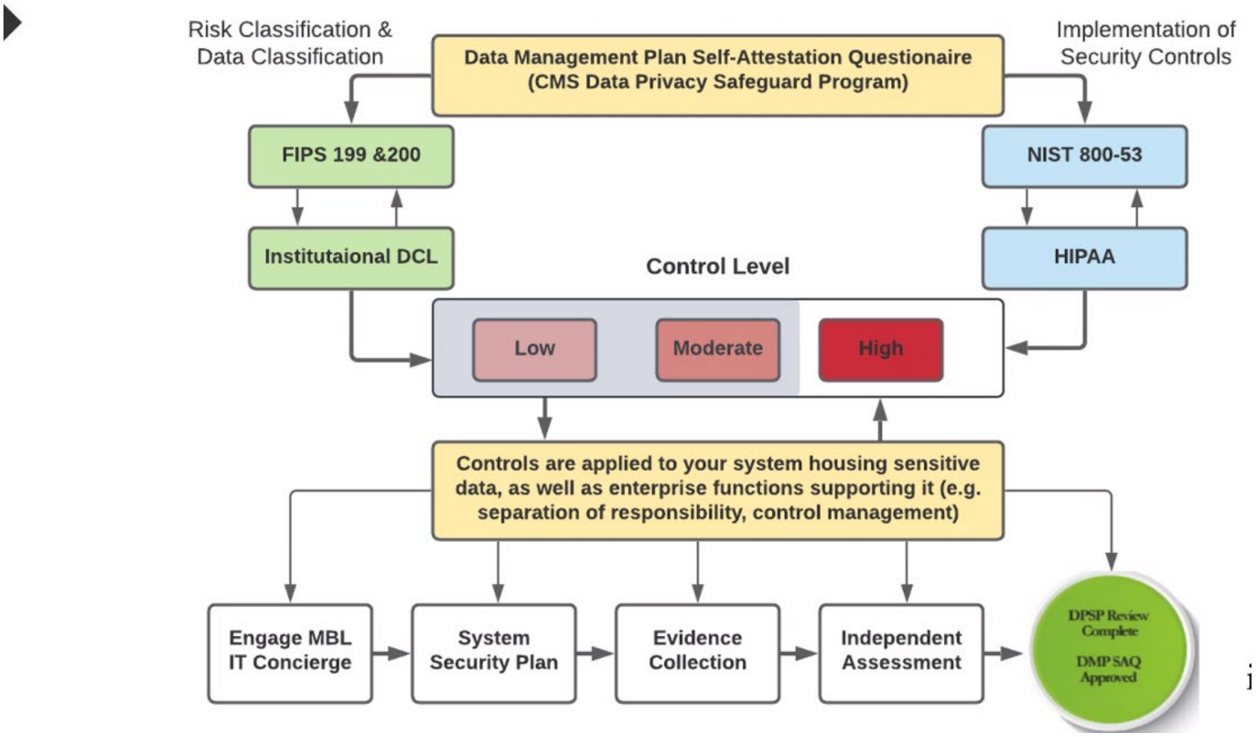
Alex Stoddard (MCW)
Q: Does 'de-identified' imply date shifting or not now?
A: The final version of the de-identified GROUSE data that researches that have access to will be date shifted. But when we received CDM data from sites, we will receive the limited version and we will do the date shifting centrally here at MU.

GROUSE Migration
-It was created in KU, and moved to MU
-Timeline- administrative work, and a lot of technical work.
-For any application that uses CMS profile you need to provide a detail data management plan
-since last year, 2020, CMS is doing a pilot procedure, where they developed a very structured way to generate this data management protocol (data management plan self-attestation questionnaire (DMP SAQ) -Use NIST SP 800-53 as a framework) a structured file to fill out and you need to provide a list of supplemental documents.
-This new data management plan is in the cloud environment in the scope.
-Another improvement that they did, new DMP, will cover all future applications using the same approve environment.

NIST SP 800-53
National institute of standards and technology, defines standards of security and privacy controls that should be used to provide environment that can support a non-regulatory government agency.

The self-attestation questionnaire process
NPL, they have concierge that we can work with.
Beside completing the questionnaire, we require you to put together, a security plan.
Need to collect evidence to show that our environment implementing all of these features by providing screenshots and scripts.
Then we provide it back to the NPL, they have a third-party auditor to look at the documents before given us greenlight.

The Self-Attestation Questionnaire Process



Risk Classification & Data Classification

Data Management Plan Self-Attestation Questionaire (CMS Data Privacy Safeguard Program)

Implementation of Security Controls

FIPS 199 &200

NIST 800-53

Institutaional DCL

Control Level

HIPAA

Low

Moderate

High

Controls are applied to your system housing sensitive data, as well as enterprise functions supporting it (e.g. separation of responsibility, control management)

Engage MBL IT Concierge

System Security Plan

Evidence Collection

Independent Assessment

DPSP Review Complete

DMP SAQ Approved

-The takeaway, even that the cloud environment is very secure and in compliance with the CMS requirements, it’s defiantly a HIPAA compliance.

DMP SAQ approval:
We work with AWS professional team, they have this program called Guardrail accelerator program, we have a contract with them and a AWS team help us work with Dev team to stand up this AWS environment that meets HIPAA and NIST SP 800-53 with minimum set of security control.

Multi -account architecture
A multi -account (best practice) was recommended by AWS for defining clear security boundaries. For the purpose of collecting all the logins to fulfill auditability requirements by CMS. The idea of multi- account is when you have something broken in one account doesn’t affect the other account.

AWS shared responsibility model
Customer is responsible for security in the cloud
AWS is responsible to ensure the security of the cloud.

Design principals
Data security
Controllability/auditability

Operational simplicity and rapid deployment – agile
Integrability
Customizability/ flexibility
Scalability -scale out and scale up
Cost-effectiveness – affordable

GROUSE data lake and Analytic Workbench
Raw data will come from NewWave GDIT, for the GPC site providers
Once the data got into the Snowflake data warehouse
Once all integration is done, data will be de-identified and accessed by the researchers.
We have built the workbench where researchers can deploy their own computing notes
And that can be directly connected through the backend database through GPC connection where researchers can perform analysis there.
The other tool that is under development is the i2b2 portal

Lav Patel (KUMC)
Q: What AWS services has been used or plan to use? Are those dedicated to GPC or shared with other AWS customers?
A: right now, we are experimenting data linkage, we haven't requested the data from the sites yet. All these data is restricted by GPC sites only.

Data provider workflow
Collect patient lists covered by CDM (SSN, Health ins claim numbers and secondary identifiers dob, gender), ADD hashed ID, then map local CDM data to CMS data. Provided to end users to use. Similar to before but have AWS.

GROUSE access workflow

- If researcher would like to use data.
- REDCap request/intake form (define funding source, study scope, covered by GROUSE IRB?, need IRB modification?) covers 3 pre-defined cohorts (Breast cancer, ALS and obesity)
- If approved – GPC DROC request
- Compliance Review (required by CMS training for study team members)
- AWS account provision for research team

Timeline

- 10/31 DUA
- NOV Instructions distributed to site for finder file generation and upload CDM data to S3 bucket
- Dec Finder files submitted
- Have 1 funded R21 – and 1 funded R01 (BP Control wants to looks at costs and utilization, not for all GPC just Utah and KUMC)
- 2 pending R01 and 1 more in preparation

- Cost recovery model allows GPC can purchase claims for 2021 – 2023. Potentially have some $s in Phase 3, but also may have some studies that can offset.
- Considering Medicaid purchase.
- On-demand data exchange in SNOWFLAKE – RW will cover later
- Expand linkage capabilities leveraging hash tokens in GROUSE – like VA data and SDOH

Q: Ryan Carnahan, options for preferred software for self service?
A: Workspace, can create your own. Like EC2 Windows