PROTOCOL TITLE:

Greater Plains Collaborative PCORnet Cohort Characterization for Breast Cancer, ALS, and Obesity

PRINCIPAL INVESTIGATOR:

Name:	Lemuel R. Waitman, PhD
Institution:	The Curators of University of Missouri
Telephone Number:	573-882-2190
Email Address:	*russ.waitman@health.missouri.edu*

POINT OF CONTACT:

Name:	Xing Song, PhD
Institution:	The Curators of University of Missouri
Telephone Number:	573-882-1352
Email Address:	*xsm7f@health.missouri.edu*

FUNDING SOURCES:

Funding	*Period*	*PI*
PCORI, Phase I-II	2014 – 2019	Lemuel R. Waitman, PhD
PCORI, Phase III	2020 – 2024	Lemuel R. Waitman, PhD
CDC/ATSDR	2022 – 2025	Xing Song, PhD; Jeffery Statland, MD

CURRENT VERSION NUMBER/DATE:

Version 2, 2023/03/30

REVISION HISTORY

***This table should only be used during submission of a Modification Application to the IRB.**

Revision #	Version Date	Summary of Changes	Consent Change?
1	2021/08/05	Initial draft	Yes
2	2023/03/30	Expand scope and formalize study oversight model	Yes

Table of Contents

Executive Summary

Background and Rationale

PCORnet was created in 2014 by the Patient Centered Outcomes Research Institute (PCORI) to further the goals of the Learning Health System and help to answer questions that are important to patient, clinician, and health system stakeholders. The Greater Plains Collaborative (GPC) is one of the 9 PCORnet Clinical Data Research Networks (CDRNs), currently consisting of 13 leading medical centers across 9 states[a] with University of Missouri being the GPC coordinating center (GPC CC).

Table 1. GPC Institutions and Catchment States

Institution Full Name	Datamart Abbreviation	Other Alias	Catchment State(s)
The Curators of University of Missouri (GPC CC)	UMO	MU	MO
Allina Health	AH	ALLINA	MN, WI
Intermountain Healthcare	IHC	IHC	UT, ID, WY, NV, MT, CO
Marshfield Clinic Research Institute	MCRF	MCRI	WI, MI
Medical College of Wisconsin	MCW	MCW	WI, MI
University of Kansas Medical Center	KUMC	KUMC	KS, MO
University of Iowa Healthcare	UI	UIOWA	IA
University of Nebraska Medical Center	UNMC	UNMC	NE
University of Texas Health Sciences Center at San Antonio	UTHSCSA	UTHSCSA	TX
University of Texas Southwestern Medical Center	UTSW	UTSW	TX
University of Texas Health Science Center at Houston	UTH	UTHOUSTON	TX
University of Utah	UU	UTAH	UT
Washington University at St. Louis	WU	WASHU	MO, IL

[a]GPC site list changes over time. The latest list can be found from the PCORnet website at https://pcornet.org/network/

Each of the PCORnet CDRNs is charged with aggregating longitudinal data on over at least 1 million individuals that is appropriate to inform all care that a patient receives, as well as creating 3 longitudinal cohorts — one rare disease, one common disease, and one to examine obesity ("3-Predefined-Cohort"). In addition, PCORI envisions PCORnet as a national resource to understand the development of health and disease in our country. In order to understand all cares a patient receives at continuum, either the CDRN's clinical delivery system needs to provide all of the care for that patient or needs to supplement delivery system data with insurance claims data that identifies care received outside of the CDRN's clinical delivery system. Throughout phase 1 (March 2014-September 2015), phase 2 (September 2015-September 2018), and phase 3 (November 2018-December 2023) of the PCORnet contract, there are contractual milestones to understand the degree to which a CDRN manages complete and comprehensive data for the patient population using the three pre-defined longitudinal cohorts as the pilot use cases. In order to assess outcomes for patients who may not remain under a single health system's care, PCORI required CDRNs to develop strategies for integrating insurance claims. Although the GPC's breadth across 9 states is advantageous for generalizing research findings, insurance carriers vary extensively; leaving integrating Center for Medicare and Medicaid Services (CMS) claims as the most consistent initial claims strategy. CMS allows qualified organizations to access their identifiable data files, also known as research identifiable file (RIF) data, for research purposes and CMS funds the Research Assistance Data Center (ResDAC) to assist investigators interested in studying CMS data. The execution of this strategy was the creation of the GPC Reusable Observable Unified Study Environment (GROUSE).

Cohort 1 - GPC-Selected Rare Disease: Amyotrophic Lateral Sclerosis (ALS)

Amyotrophic Lateral Sclerosis (ALS) is a multisystem neurodegenerative condition with predominant motor system involvement. It is an invariably fatal neurodegenerative disorder, with a median survival of 3-4 years from symptom onset.[1] ALS is rare. The worldwide all-age prevalence was 4.5 (95% CI 4.1 – 5.0) per 100,000 people, and incidence was 0.78 (95% CI 0.71 – 0.86) per 100,000 person-years.[2] In United States, 2016 estimates from the National ALS Registry suggested a prevalence rates of ALS at 5.2 (95% CI 5.1 – 5.3) per 100,000 of the population.[3] ALS has a high disease burden. According to the 2016 global disease burden Motor Neuron Disease Collaborators (GBD-MNDC) study, 330,918 individuals had a motor neuron disease, causing 926,090 disability-adjusted life-years (DALYs) and 34,325 (95% bootstrapped confidence interval [CI] 33,051 – 35,364) deaths in a single year.[2] ALS incidence is rising. While no unifying pathogenesis has been described across the entire spectrum of ALS phenotypes, the incidence of the condition is

projected to rise from 222,801 in 2015 to 376,674 in 2040, representing an increase of 69% in the next couple of decades.[4] Besides confirmed genetic factors, speculated occupational and environmental risk factors that have been reported, large-scale epidemiology study also suggests "ALS clusters" related to socio-economic status (SES) and geographical location. Nonetheless, at this time the best biomarker that can be used in ALS research is the presence or absence of one of these known genetic defects. We will capture this information, when known, as part of the GPC activities with respect to ALS.

As of early 2023, the only pharmacological treatments that have shown to slow ALS are Riluzole, Endaravone and AMX0035.[5,6] However, there are a number of symptomatic therapies available for ALS patients. The American Academy of Neurology has published practice parameters guidelines[7] and the ALS clinical and research community has established a number of clinical measurement tools to assess outcomes that can be used by the GPC to do comparative effectiveness studies on the symptomatic management of ALS, as there are more than one standard of care approaches to symptom management in ALS. These include the management of pseudobulbar affect (emotional lability), excessive drooling, and cognitive function. We have identified a number of measurement scales that we can use in comparative effectiveness studies to assess these symptoms. These include: the ALSFRS,[8] the ALS Cognitive Behavior Screen,[9] the ALS quality of life measure,[10] the Center for Neurologic Study-Lability Scale,[11] and the Drooling Scale.[12,13] Three of these measures (the ALSFRS, the ALS Cognitive Behavior Screen and the Lability Scale) are part of the NIH Common Data Element project for ALS[14] which also contains demographics, symptom-sign- diagnosis criteria, neuromuscular and physical examination, and vital sign data.

Cohort 2 - GPC-Selected Common Disease: Breast Cancer

Breast cancer is primarily a female disease (1% of all breast cancers are among men). Based on NCI Surveillance Epidemiology and End Results Program data from 2008-2010, it is estimated that 1 in 8 women born today will be diagnosed with breast cancer during their lifetime. On January 1, 2010, in the United States there were approximately 2,829,041 women (1.2% of all adult women) who had a history of cancer of the breast. In 2013 alone it is estimated that 232,340 women will be diagnosed with, and 39,620 women will die from, breast cancer. Median age at diagnosis is 61 years and 93% of patients are diagnosed when their disease is localized or regional. Five-year relative survival is high (98.6%) for the 61% of cancers that are confined to the primary site, lower (84.4%) for the 32% that have spread to regional lymph nodes at the time of diagnosis, and poor (24.3%) for metastatic disease.[15] Breast cancer incidence has increased with time, historically attributed to changes in known reproductive risk factors (number of children, age at first birth) and more recently to the availability of mammography screening and growing obesity and post-menopausal estrogen use (until 2002 when use dropped precipitously). In spite of comparable trends in incidence rates, improvement in survival rates among African American women has not kept pace with that among white women. Tumors are larger and the proportion diagnosed when metastatic is higher among African American women. Survival is also lower among young women (< 40 years) whose tumors may be more aggressive.[16]

These statistics belie the heterogeneity of breast cancers and the poor survival and limited treatment options for important clinical subgroups such as patients with "triple-negative" cancers.[17] Breast cancer is one of a few examples where "personalized therapy" has been realized, in which therapies target specific molecular subtypes of the disease. However, access to such therapies is expected to vary along the same socioeconomic lines as traditional health disparities and patient-centered studies to eliminate the quality chasm are needed. Patient-centered comparative effectiveness studies nested in rich data systems also are needed to answer questions that patients and their physicians need to know but which are not the typical questions addressed in phase 1-3 cooperative group trials. When mature, the PCORI national network should greatly enhance and complement existing cooperative group trial mechanisms—for example, to assist in efficient recruitment to reduce trial costs and ascertain sufficient cohort sizes. We believe the GPC and the national CDRN-PPRN network will provide the breakthrough needed to accelerate research progress by providing the infrastructure needed to access and work with patients in CER and other research on breast cancer and its treatment.

Cohort 3 - PCORI Mandated Condition: Obesity

Obesity continues to receive considerable attention in both the scientific and lay literature due to the rapid rise in the prevalence of obesity over the past two decades. The general consensus is that a major cause of this is an environment that promotes caloric intake and discourages caloric expenditure. Most of the short- and long-term consequences of obesity for the individual are well-known and well-described, ranging from orthopedic disorders to diabetes and cardiovascular disease. The economic consequences of obesity-related illness are rapidly approaching $200 billion, representing about 25% of annual medical spending in the US.[18] The indirect and mortality costs are 2-3 folds greater than the direct costs.

Childhood obesity,[19] in particular, has increased significantly in the past two decades. For example, in the early 1990s (1988-1994, NHANES III)) prevalence of obesity in adolescent boys (12-19 years) was 11.3%. By 2009-2010 this increased to 19.6%. An even more dramatic rise was seen among young boys (2-5 years), with prevalence rates more than doubling (6.2% to 14.4%). Childhood obesity appears to disproportionately affect certain racial and ethnic groups. For example, the prevalence of obesity among non-Hispanic black girls is 24.8% compared to 14.7% for non-Hispanic white girls. The resulting health and economic devastation over the next few decades is unimaginable, especially since many of the "long-term consequences of obesity" (e.g., type 2 diabetes and cardiovascular disease) are now being seen in adolescents.[20] Even in the pediatric population, analyses suggest that both the number of hospitalizations and the costs of hospitalizations directly related to obesity are rising rapidly.

The states in which our GPC sites are located contain about 15% of the adults and 17% of the children in the US. Collectively, these states have a slightly lower prevalence of overweight (22.4% vs. 23.3%) and obesity (29.0% vs. 35.9%) in adults when compared to the US as a whole, and a higher prevalence of overweight/obesity (28.0% vs. 18.0%, separate data not available) in children, compared to the US as a whole. Ethnically, the GPC states have a higher percentage of Hispanics (23.4%) and a lower percentage of Blacks (9.7%) than the US as a whole (16.9% Hispanic, 12.1% Black.)

The impact that childhood obesity is having on outcomes has been examined in some disease-specific populations. For example, there is an increasing prevalence of overt diabetes during induction therapy or as a long-term consequence of acute lymphoblastic leukemia (the most common childhood malignancy). Patient-centered comparative effectiveness studies nested in rich data systems are needed to answer questions that impact health care systems, influence provider behaviors, and alter the natural trajectory of obesity and its consequences, such as diabetes, cardiovascular disease, and orthopedic disorders.

Data Quality and Care Quality: Capture of Complete and Comprehensive Clinical Information

Mirroring national trends, the timing of EHR deployment varies across GPC sites and, in some cases within individual sites. For example, at the University of Kansas, Epic was implemented in inpatient units in December 2007, while outpatient implementation occurred in a staggered manner between 2010 and 2012. It is particularly noteworthy that at some sites, data repository records extend for decades, contributing to very large denominators. For example, the University of Wisconsin has over a century of certain data sources and the University of Kansas, University of Iowa and Marshfield Clinic have long-standing tumor registries. Collectively, however, the GPC can address a wide range of observational and longitudinal studies, despite variations in implementation across sites.

We are excited about the many strengths and diversity of the populations served by our GPC sites; however, we recognize limitations exist in most health systems for comprehensive, complete, and longitudinal data capture, and the GPC is no exception. First, patients exercise choice and often receive care from multiple systems. For example, even within the VA healthcare system, the nation's largest integrated delivery system, more than 50% of patients receiving VA care in any one year also receive some services from the private sector. Similarly, it has been well chronicled that patients who are insured through HMOs and Medicare Advantage plans often receive services through the VA as well. Second, patients change providers over time as a result of changes in their health care benefits. Third, patients are referred out of smaller systems to receive specialized care only available at tertiary centers. Fourth, while the GPC sites collectively have very large (and growing) primary care populations who receive the majority of their care at the GPC sites, GPC sites provide a considerable amount of specialty care to large numbers of patients for whom the site may not have routine, longitudinal primary care data. In this context, payer and health plan claims data can provide information on out of system health care utilization. Our funded contract for PCORI proposed supplementing our data repositories with Medicare/Medicaid claims data from CMS through ResDAC to meet our contractual obligations to increase the overall data completeness.

Goals and Objectives

Within the context of GPC and PCORnet, our study focuses on the following overarching goals:

(1) To understand the development, treatment, progression and consequences of ALS, breast cancer, and healthy vs. unhealthy (overweight and obesity) weight.

(2) To evaluate and enhance data quality derived from electronic health records (EHR) and claims, as well as through integration of other 3rd party data resources.

(3) To examine care disparities at individual, community and institution level (e.g., comparisons on access to care among Medicare/Medicaid-insured, commercially-insured and uninsured population) and evaluate generalizability of pragmatic interventions.
(4) To evaluate healthcare utilization and the economic impact Evaluate healthcare utilization and the economic impact of multiple acute and chronic conditions and identify "hotspotting" areas to better inform policies for lowering costs and patients' financial burden.
(5) To serve as a greater national resource to understand the development, treatment, progression, and consequences of acute and chronic disease cared for within the United States healthcare system and in support of quality care for the conditions championed by the Patient Powered Research Networks in PCORnet as well as the other conditions studied by our peer CDRNs.

To support these aims from a methodological basis, we seek to expand the data completeness of our patients' health care processes and outcomes and understand the information gain for our complete and comprehensive population by comparing correlations between the Medicare and Medicaid claims data with the data in our CDRN that includes the electronic health record and billing data from each of our component health systems, clinical registry data (e.g., hospital tumor registries), private payer claims data, and patient-reported outcomes, as available, and work with our GPC and CDRN investigators to answer specific cohort questions to achieve our overarching aims.

Our first methodological hypothesis is that when our health systems function as tertiary or quaternary care facilities, they will hold very comprehensive (detailed, multifactorial signals, structured and un-structured) indicators of health and care processes but as patients return to their primary care and especially rural environments, the Medicare and Medicaid claims data will provide valuable signals of follow up care processes and outcomes. For example, a breast cancer patient living in rural Kansas may have come to KUMC for diagnosis and returned to their local providers for treatment. In another scenario, a patient may be referred by their primary care physicians to our health systems for procedures and specialty care. Thus, in this context, we seek to **measure the contribution of various data provenances to information completeness in capturing patients' care pathways and health trajectory**.

Second, there is considerable variability in the mix of primary versus specialty care within our health systems and in comparison to the larger United States healthcare system composed of community hospitals and providers. The degree to which analyses based upon the cohorts contained in the GPC reflect the larger populations within our states is not well characterized and our hypothesis is that it may be condition specific. For example, ALS patients are usually seen at tertiary care centers such as Amyotrophic Lateral Sclerosis Association (ALSA) or Muscular Dystrophy Association (MDA) sponsored clinics due to their need for a knowledgeable and experienced multidisciplinary health care team including Physical Therapists, Occupational Therapists, Respiratory Therapists, Speech Therapists, Social Workers, Nurses, Equipment Vendors, and Dieticians. Since each GPC site has an ALS clinic sponsored by ALSA or MDA or both, we should cover most of the ALS patient population in our states; however, whether our patient population represents the larger population in the regions is still in question. Similarly for other diseases such as breast cancer and obesity, we may only cover a small fraction of the patient population in our regions. Thus, our second objective is to **calculate the distributions of our three conditions and their treatment patterns for our cohorts within the GPC relative to their occurrence in our larger state regions**.

Finally, during the tertiary care periods, understanding the correlation between EHR, billing and claims data is not well described for a diverse set of health systems encompassed by the GPC. Most of the risk adjustment and health services research has rested on standardized claims data from sources such as CMS via ResDAC. Our ability to use rich and more current clinical and administrative data to drive trial recruitment and observation for PCORnet will hinge on the ability to **quality control these new data sources and quantify the relative support of EHR and billing systems-based computable phenotypes versus established claims-based models**. This will be accomplished for diagnoses, procedures, hospitalizations/home care/ambulatory visits, medications, and provider characteristics between EHR and CMS claims. A specific example will be to compare diagnostic code assignments in claims versus EHR and billing data for cohort characterization, determining how encounters matchup between claims and EHR and how accurate and timely EHR and billing systems-based computable phenotyping is against claims. Some specific research questions we are looking to investigate for our three selected conditions are listed in Table 2.

Table 2. Sample of Specific Research Questions

Amyotrophic Lateral Sclerosis (ALS)

- Compare the effectiveness of symptomatic management of ALS such as the management of pseudobulbar affect (emotional lability), excessive drooling, and cognitive function
- Compare the assessment of the ALS symptoms using measurement scales such as the ALSFRS, the ALS Cognitive Behavior Screen, the ALS quality of life measure, the Center for Neurologic Study-Lability Scale, and the Drooling Scale
- How often do ALS patients receive their follow-up care through our clinics versus other providers?
- Compare treatments and outcomes of ALS patients from diagnosis to death for the overall CMS population versus those managed by our clinics
- How many ALS patients are on FDA-approved medications (Riluzole, Endaravone, AMX0035), versus symptomatic medications (e.g., anticholinergic for drooling)?
- Compare the basic diagnoses of complications such as DVT/PE (deep vein thrombosis/pulmonary embolism) and aspiration pneumonia in the EHR, billing, and claims across GPC sites
- How do code assignments of PEG tube placement and BIPAP in EHR compare against CMS claims?
- How do the basic demographics of our ALS patients compare to the overall state population?
- Development and Validation of diagnostic and/or prognostic models of outcome
- Mediation analyses of biological, cultural and environmental factors that affect treatment response or course of disease
- Discovery or validation of multi-domain clinical and/or biological measures for diagnosis, prognosis and/or treatment response using existing genetic and biological samples along with clinical and physiological assessments
- Extended characterization or validation of natural history disease course and novel methods for improving patient stratification
- Projects which include a focus on health disparities and inequities in neurological disease, healthcare, and health outcomes in disparate populations, including racial and ethnic minorities, the geographically disadvantaged, sex and gender minorities, and others who have been historically underserved, socioeconomically disadvantaged, marginalized, or adversely affected by persistent inequality.

Breast Cancer

- Are there serum markers or radiologic findings (breast density) in high risk patients that predict for development of disease?
- Compare the effectiveness of using Magnetic Resonance Imaging (MRI) or mammography in young patients and women who have dense breasts to improve health outcomes (i.e., survival, local recurrence, mortality)
- Compare different approaches to manage ductal carcinoma in situ (DCIS)
- Compare the safety and effectiveness of therapies for reducing risk of breast cancer among high risk patients
- Compare the disease-free survival and health-related quality of life for patients receiving molecularly-guided therapy versus usual care
- Compare the adherence, overall survival, and health-related quality of life for patients receiving various diet/exercise interventions
- Compare effects of survivorship care planning models on cancer follow-up care, preventive services, health promotion behaviors, and patient-reported outcomes
- Analyze effect of genetic testing on treatments and patient reported outcomes
- Determine predictors and outcomes of breast reconstruction options including complications, patient reported outcomes, and oncologic safety
- Projects which include a focus on health disparities and inequities in neurological disease, healthcare, and health outcomes in disparate populations, including racial and ethnic minorities, the geographically disadvantaged, sex and gender minorities, and others who have been historically underserved, socioeconomically disadvantaged, marginalized, or adversely affected by persistent inequality.

Healthy vs. Unhealthy Weight

- Compare the effectiveness or side-effects of medication dosing based on weight versus lean body mass versus body surface area versus BMI
- Examine the effectiveness of implementation of pediatric or adult guidelines by providers in reference to particular diseases (e.g., blood pressure screening) and variability across providers
- Explore the impact of obesity on specific outcomes in the hospital (e.g., risk of wound infection in obese and lean individuals undergoing Caesarian section) or in the outpatient setting (e.g., frequency of ER visits for asthma in lean and obese children)
- Compare different approaches to management (lifestyle, telemedicine, surgery, medication) of obesity in different populations (children, young adults, elderly)
- Compare the safety and effectiveness of interventions for obesity in populations of various ages and track short and long term outcomes (e.g., gastric bypass at 16-20 years of age versus 26-30 years of age)
- Compare the health-related quality of life for patients receiving surgery or medication versus lifestyle interventions
- Identify prevalence of polymorphisms and abnormalities in typical obese, overweight and normal populations
- Explore the impact of obesity on health care costs in specific populations (e.g., duration of hospitalization after coronary artery bypass)
- Compare the effectiveness of including weight management as a routine intervention with orthopedic injuries involving lower limbs
- Projects which include a focus on health disparities and inequities in neurological disease, healthcare, and health outcomes in disparate populations, including racial and ethnic minorities, the geographically disadvantaged, sex and gender minorities, and others who have been historically underserved, socioeconomically disadvantaged, marginalized, or adversely affected by persistent inequality.

Data Quality and Care Quality

- Evaluate and enhance data quality and completeness derived from electronic health records (EHR) and claims, as well as integration with other 3rd party data resources.
- Investigate treatment patterns (e.g., follow-up care occurrence, long term care placement timings, and hospitalization details) and outcomes of patients with acute or chronic conditions, especially those with complex management needs and heterogenous treatment responses (not limited to the 3 pre-defined cohorts).
- Examine pathway to diagnosis and outcomes of patients with acute or chronic conditions, especially those with complex onset representations and susceptible to delayed diagnosis (not limited to the 3 pre-defined cohorts).
- Evaluate healthcare utilization and the economic impact of multiple acute and chronic conditions and identify "hotspotting" areas to better inform policies for lowering costs and patients' financial burden.
- Examine care disparities at individual, community and institution level (e.g., comparisons on access to care among Medicare/Medicaid-insured, commercially-insured and uninsured population) and evaluate generalizability of pragmatic interventions among patients of different protected classes based on demographic and socio-economic-status (not limited to the 3 pre-defined cohorts).

Data Sources, Linkages and Justification

Electronic Health Records (EHR)

All GPC sites have aggregated longitudinal EHR data using the PCORnet common data model (CDM) ensuring interoperability of data across multiple institutions. The current version of PCORnet CDM integrated on GROUSE contains curated clinical data over 2010-2022 from all 13 GPC sites, which includes patient demographics, vital signs, smoking history, labs mapped to LOINC codes, prescribing medication mapped to RXNORMs, dispensing medication mapped to NDC, procedures mapped to CPT/HCPCS and ICD10, diagnoses mapped to ICD9 and ICD10 codes.

Hospital Tumor Registry

Since its beginning, the GPC has prioritized inclusion of tumor registry data in each institution's datamart leading to publications and provided motivation for sites to invest time into populating a standardized tumor table. The Cancer Collaborative Research Group (CRG) was funded by PCORI between 2017 and 2019 to promote multi-network cancer research and develop data science resources. One resource that emerged from this project was a set of specifications for a tumor table that supplements the PCORnet Common Data Model with cancer-specific data. This table primarily holds data from hospital tumor registries that are formatted according to standards published by the North American Association of Certified Cancer Registrars (NAACCR). All hospitals that are accredited by the American College of Surgeons Commission on Cancer employ trained registrars to abstract medical record data according to these specifications. Structured fields for demographic, clinical, and treatment observations are included, and the data are considered to be high quality. The tumor table documentation developed by the Cancer CRG includes specifications for data formats, quality checks, and relationships with other CDM tables. This standardization allows linkages between NAACCR data and the other CDM tables. It also allows queries of the NAACCR data to be quickly deployed across the network. Most GPC sites have successfully implemented the CDM Tumor table derived from site-level hospital NAACCR, which will be a critical data source to support cancer research.

Medicare and Medicaid Claims (CMS)

Currently, the GPC coordinating center (GPC CC) have purchased multi-year, multi-state Medicaid and Medicare RIF files from CMS chronic condition warehouse (CCW) and will potentially acquire claims data for subsequent years pending additional federal grant funding using PCORnet and GPC. Current Medicaid and Medicare RIF files are listed below:

- **Medicare Enrollment and Beneficiary-Level (MBSF) file [2011 - 2020]**: MBSF file, or denominator file, is created annually and contains demographic entitlement and enrollment data for beneficiaries who: 1) were part of the user-requested sample; 2) were documented as being alive for some part of the reference year; and, 3) were enrolled in the Medicare program during the file's reference year.

- **Medicare Provider Analysis and Review (MedPAR) [2011 - 2020]**: The MedPAR file includes all Part A short stay, long stay, and skilled nursing facility (SNF) bills for each calendar year. MedPAR contains one summarized record per admission. Each record includes up to 25 diagnoses (ICD9/ICD10) and 25 procedures ((ICD9/ICD10) provided during the hospitalization.

- **Outpatient Claims [2011 - 2020]**: The outpatient file contains Part B claims for 100 percent for each calendar year from institutional outpatient providers. Examples of institutional outpatient providers include hospital outpatient departments, rural health clinics, renal dialysis facilities, outpatient rehabilitation facilities, comprehensive outpatient rehabilitation facilities, community mental health centers. In and out surgeries performed in a hospital will be in the hospital outpatient file, while bills for surgeries performed in freestanding surgical centers appear in the carrier claims, not in the outpatient file.

- **Carrier Claims (NCH) [2011 - 2020]**: Since 1991, the Center for Medicare & Medicaid Services (CMS) has collected physician/supplier (Part B) bills for 100 percent of all claims. These bills, known as the National Claims History (NCH) records, are largely from physicians although the file also includes claims from other non-institutional providers such as physician assistants, clinical social workers, nurse practitioners, independent clinical laboratories, ambulance providers, and stand-alone ambulatory surgical centers. The claims are processed by carriers working under contract to CMS. Each carrier claim must include a Health Care Procedure Classification Code (HCPCS) to describe the nature of the billed service. The HCPCS is composed primarily of CPT-4 codes developed by the American Medical Association, with additional codes specific to CMS. Each HCPCS code on the carrier bill must be accompanied by a diagnosis code (ICD9, ICD10), providing a reason for the service. In addition, each bill has the fields for the dates of service, reimbursement amount, encrypted provider numbers (e.g., UPIN, NPI), and beneficiary demographic data.

- **Durable Medical Equipment Claims (DME) [2011 - 2020]**: The Durable Medical Equipment (DME) file contains fee-for-service claims submitted by Durable Medical Equipment suppliers to the DME Medicare Administrative Contractor (MAC).

- **Part-D Drug Event and Drug Characteristic files, [2011 - 2020]**: When a Medicare beneficiary with Part D coverage fills a prescription, the prescription drug plan submits a record to CMS. The PDE file includes all transactions covered by the Medicare prescription drug plan for both Prescription Drug Plans (PDPs) and Medicare Advantage Prescription Drug Plans (MA-PDs).

- **Part-C Medicare Advantage Encounter files, [2018 - 2020]:** The reason for requesting Medicare Advantage claims data is import in further improving data continuum and data representativeness, especially for capturing clinical activities happening outside the healthcare systems of our GPC network. From 2011 to 2020, the Medicare Advantage penetration rate has been increased from 25% to 40%, and such increasing trend was shown to be consistent across states. Most of our current study requires an exclusion of patients who were enrolled in Part C, which not only significantly reduced the statistical power but may also introduce potential biases against beneficiaries with more severe conditions.

- **Medicaid MAX files, [2011 – 2012]**: the obsolete Medicaid MAX files including Person Summary, Inpatient files, Prescription Drug files, Other Therapy files, and Long-Term-Care files.

For better interoperability, we have transformed most parts of the Medicare RIF files (only FFS claims) into PCORnet CDM format except for charges and costs. We are currently working on developing transformation codes for Medicare Advantage data files, Medicaid TAF files, as well as designing and developing the process of augmenting PCORnet CDM to contain economic data transformed from Medicare/Medicaid charges and costs data.

Linkage between EHR and CMS data

HealthAPT is a Joint Venture (JV) business entity formed under a U.S. Small Business Administration Mentor-Protégé Program. HealthAPT is established through the partnership of NewWave Technologies Inc. and General Dynamics Information Technology (GDIT), both deeply experience and dynamic CMS contractors. HealthAPT operates and maintains the Chronic Condition Warehouse/Virtual Research Data Center under a contract with CMS. All data files provided to HealthAPT, including finder files, are protected by the privacy and security language contained in the official CMS contract. The contract contains the required Health Insurance Portability and Accountability Act business associate language and HealthAPT has a completed and signed Data Use Agreement (DUA) on file with CMS. As a CMS contractor, HealthAPT acts on behalf of CMS and is permitted to receive identifiable data (finder files) from researchers. All data received by HealthAPT can and will only be used to fulfill researcher's data request.

As shown in Figure 1, individual GPC sites will first send the NewWave-GDIT (Research Data Distribution Center under contract with CMS) encrypted beneficiary identifiers (e.g., SSN, HICs, name, DOB) along with a hashed IDs generated from the hashing algorithm (reference figure below). After receiving beneficiary information from each GPC sites, NewWave-GDIT will generate the finder files that contain mappings between the hashed IDs and CMS IDs and send the crosswalk files and CMS data back to MU (not individual GPC sites). Finally, MU will integrate the crosswalk and CMS data with individual site EHR data (limited data set) through linking the hashed IDs on their secure virtual private cloud. This will allow MU to achieve record linkage of patients' EHR and claims data without

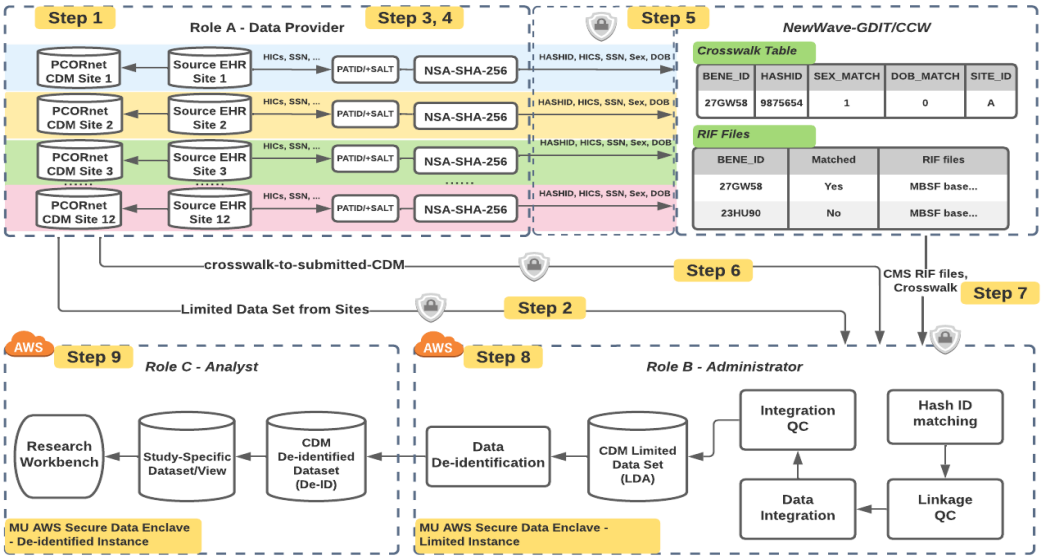

Figure 1. *EHR-CMS Data Linkage. NSA-SHA-256 stands for National Security Agency Security Hash Algorithm, which is set of cryptographic hash function approved by NIST for encrypting identifiers. SALT stands for a set of random bits added to each identifier before applying the SHA-256 hashing algorithm. QC*

obtaining or retaining actual patient level PII from individual GPC sites. At last, the merged dataset will be de-identified and made available on a separate virtual private cloud to project team members for running analyses or using the i2b2 client.

In order to maintain an up-to-date linkage between the CMS and site EHR data, we seek to refresh the patient crosswalk file (i.e., a mapping between patients' local identifier to CMS beneficiary ID) on an annual basis until at least one year after the funding period (2025) if not further extended. This activity will involve participating sites (including MU) sending PHI information (SSN, HIC, DOB) to CMS's data warehouse contractor, New-Wave/GDIT. After each refresh, all the finder files GDIT receives are kept locked in a secure location after the data is uploaded to our data processing system. The media is kept for 90 days and then destroyed using commercial media destruction methods.

Explain how the data to be used are reasonable and necessary to conduct research

State-level CMS claims: This cohort will include only individuals residing in GPC regions (i.e., GPC Catchment States) supported by the GPC institutions NIH Clinical and Translational Science Award/Great Plains IDEA – Clinical and Translational Research (Table 1). The geographic region described contains approximately 25 million beneficiaries. Because people may move between health systems, it is necessary to cover the regions, rather than the individuals, covered by the health systems. Therefore, to accomplish the second aim in comparing our three studied conditions and their treatment patterns against the larger population within the states, it is necessary to obtain claims data for all patients residing in the eleven states.

Linkage to site EHR data: The purpose of medical claims data is to support payment for care, which may not accurately reflect a particular disease. Integration and linkage with site EHR data will provide not only more accurate but more granular clinical observables beyond what are available in claims data (e.g., lab values, vital signs, flowsheet documentations, physician notes). For example, body mass index information is only available through EHR, which is the most reliable marker to identify overweight and obesity cases.

Integration of cancer registry: The curated and cancer-specific data elements available in site-level hospital cancer registry provide critical information to address proposed research questions related to breast cancer.

Preservation of non-CMS and non-EHR population: Although many research questions can be answered by the crosswalk population (i.e., Medicare or Medicaid beneficiaries who are also included in the site EHR data), it is critical to retain the non-CMS (or EHR-only patients who were only observed in the EHR data) cohort and non-EHR cohort (or CMS-only beneficiaries who were only observed in the Medicare and Medicaid claims data).

Linkage to National Plan and Provider Enumeration System

We will link national provider identifier (NPI) to the CMS national plan and provider enumeration system (NPPES), for quality assurance of the claims provider versus the providers identified in the electronic medical record for the encounter, as well as comparing care pathways and accessibility within and external to our participating GPC institutions.

Geocoding and Linkage to Public Microdata Files

All Medicare and Medicaid beneficiaries with valid Zip+4 records will be geocoded and linked with the following publicly available data or microdata files (Table 3), which usually are aggregated values at a geographical level (e.g., census block group, census tract, county, 5-digit Zip).

Table 3. Public Data Domains

Data Domain/Source	Geographic Unit	Data Provider	Potential # of Markers
Socio-marker			
Distressed Communities Index	Zip code	EIG	7
Medically Underserved Area	County	HRSA	1
Social Vulnerability Indices	Census tract, county	CDC, UCA	2
Food Access Research Atlas	Census tract	ERS	140
Rural-Urban Commuting Area codes	Census tract	ERS	2
Social Deprivation Index	Census tract	RGC	1
Area Deprivation Index	Census block group	HRSA	1
American Community Survey (ACS)	Census block group	CB	3,263

Geo-marker			
Air Quality System Database (AQS)	County	EPA	555
National Water Information System	Zip+4 code	USGS	99,998
National Hydrography Dataset Plus Data	Zip+4 code	EPA	153
National Oceanic and Atmospheric Administration Datasets	Zip+4 code	NCEI	219
Soil Survey Geographic Database	Zip+4 code	NCSS	600

ACS – American Community Survey; EIG – Economic Innovation Group; RGC- Robert Graham Center; ERS – Economic Research Service; CB – Census Bureau; HRSA – Health Resources and Services Administration; CDC – Centers for Disease Control and Prevention; EPA – US Environmental Protection Agency; CMS – Center of Medicare and Medicaid Services; NHANES – National Health and Nutrition Examination Survey; USGS – United States Geological Survey; NCEI – National Centers for Environmental Information; NCSS – Natural Resources Conservation Service Soils

Other Data Resources via Datavant Linkage

The Datavant software solution enables Privacy-Preserving-Record Linkage (PPRL) through the use of de-identified, encrypted tokens that can be linked across data sources. With Datavant De-ID software, the underlying PHIs are irreversibly hashed (i.e., cannot regenerate the PII from the hash value) into a series of Master Tokens using the Datavant Master Seed. The key-based hashing process means that the same PII processed at one site will result in different hashes from those produced from another site. In order to match tokens from different sites, the site-specific Master Tokens are then transformed into Transit tokens that are specifically directed at a third-party site using a site-specific key. This means that if the transit tokens are accidentally sent to the wrong location, the transit tokens cannot be processed by the incorrect site. The receiving site can then transform the transit tokens from different sites into tokens that can be used to match the same patients across different sites. The same PHI always generates the same set of Master Tokens, but it is never present in any output or log stream from the De-ID software. Only the site-specific encryption tokens are written to the output file. All GPC sites have an existing Site License Agreement (SLA) in place with Datavant and already have access to the Datavant De-ID software. Datavant-based linkage will be study- and provenance-specific. Upon approval by GPC study oversight committee, a study-specific Order Form (usually a NO COST, NO SIGNATURE form) will need to be provided by Datavant to supplement the extant SLA.

Subject Selection and Withdrawal

Inclusion Criteria

We will include all observable patients included within the GPC network as well as beneficiaries of Medicare and/or Medicaid who resides in the catchment states of GPC institutions.

Exclusion Criteria

To minimize selection bias, we will not apply any general exclusion criteria to the overall GROUSE cohort. Site EHR data should be completely excluded from a particular study if they opt-out from the study.

Key Considerations for Cohort Selection

In general, selection bias should at least be considered based on data coverage. As shown in Figure 2, population A (*the crosswalk cohort*) can be observed with the most complete data but duplications over multiple source; population B (*the non-EHR or CMS-only cohort*) would have better data continuum compared to population C (*the non-CMS or EHR-only cohort*); population C would have more granular data capture especially on specialty cares or hospitalization episodes; population $A + C$ represents the *full-EHR cohort*; population $A + B$ represents the *full-CMS cohort*. Geocoded data might be only available for population $A + B$ (the *full-CMS cohort*). It is worth noting that sites may choose to opt out from studies that involve population C (see Study and Data Oversight section for more details).

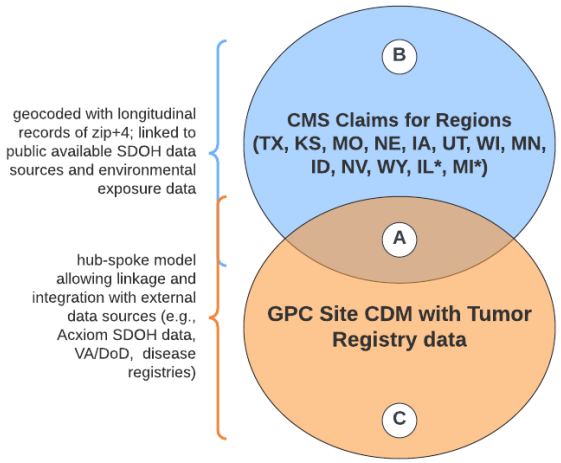

Figure 2. *Selection Bias relevant to Data Provenance*

Risk and Benefit

Potential Risks

Minimal risk is considered for this study. While unique identification numbers or personal health information (PHI) such as surnames, given names, date of birth, and address would make record linkage straightforward, distribution of such information is however restricted due to privacy concerns. Since patient's trust is with their home health systems and academic medical centers, we seek to avoid transmitting PHI through the GPC network infrastructure (i.e., from GPC site to GPC CC). The CMS contractor, HealthAPT, will serve as the intermediary for brokering file linkage between site EHR to CMS.

Only limited datasets (LDS) are allowed to directly transfer from different data providers (e.g., GPC sites, CMS CCW) to GPC CC in a secured fashion (e.g., encryption-in-transit). Raw data are stored in a NIST-800-53-compliant and HIPAA-compliant database following encryption-at-rest and password-protected protocol (mor details described in the *Data Management Plan* section). Only fully de-identified dataset with approved cohort and data elements (e.g., exclude opt-out sites) will be released for research purpose and accessible to approved and trained study team members.

Potential Benefits

There are no known potential benefits to this study to subjects, but there is a potential benefit to scientific discoveries and public health. Study findings will contribute to real world evidence for better understanding of diseases, as well as inform better trial design and best practices of utilizing real world data. In addition, the integrated database will serve as a greater national resource to understand the development, treatment, progression, and consequences of acute and chronic disease cared for within the United States healthcare system and in support of quality care for the conditions championed by the PPRNs in PCORnet as well as the other conditions studied by our peer CDRNs.

Study and Data Oversight

Multi-Stakeholders and Governance Structure

Study and data oversight is jointly provided by Research Opportunity Committee (ROA), Data Request Oversight Committee (DROC), and Study Oversight Committee (SOC). All GPC-related data requests are required to be reviewed by DROC and SOC. If data request is determined to be beyond the keystone DUA, additional review will be required by ROA and a separate DUA and IRB may be required.

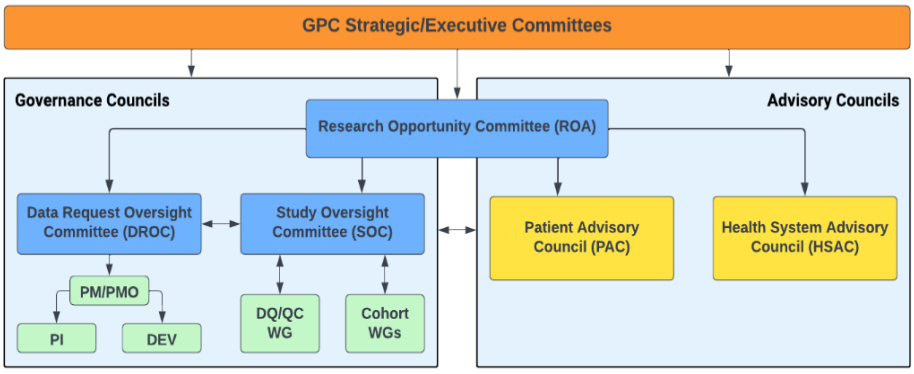

Figure 3. *Stakeholders and GPC Governance Structure*

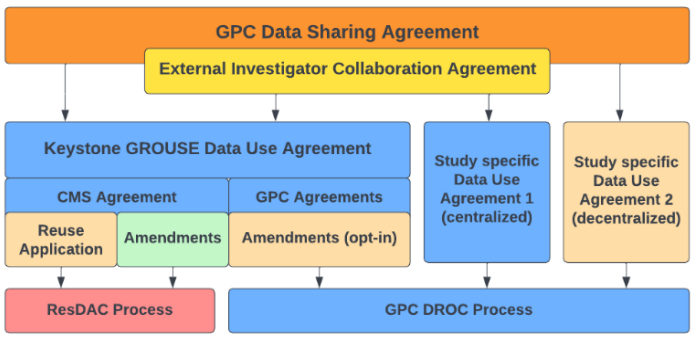

Figure 4. *Consortium Regulatory Model*

Consortium Regulatory Model

The *Keystone GROUSE DUA* consists of the master CMS DUA (with amendments) and GPC DUA (with amendments), which currently only covers studies within the scope of the three GPC-selected cohorts (ALS, breast cancer, healthy vs. unhealthy weight) on *the crosswalk cohort*. Studies with extensive EHR data access should be covered by appropriate GPC DUA amendments (opt-in) and reviewed by DROC and endorsed by SOC. Studies with different scope would require either a full reuse application to CMS and/or a study-specific DUA with GPC sites.

Data Management Plan

GPC Cloud-based Data Enclave

Figure 5 illustrates system architecture of the GROUSE environment, which is composed of a data lake, a data warehouse and analytic workbenches. A) "Data Lake": data (including GDIT physical media) are loaded into secure S3 buckets via Secure Shell File Transfer Protocol (SFTP) or Transport Layer Security (TLS) 1.2 Protocol. B) "Data Warehouse": data is extracted and loaded into Snowflake for data transformation into the PCORnet Common Data Model and de-identification. C) "Analytic Workbench": to minimize the burden on researchers of learning to navigate the cloud environment, we adopted an AWS solution —service workbench, where approved users can self-service to deploy either Windows or Linux analytic "workspaces" of multiple analytical applications (e.g., R, Python, SAS) and varying computing power based upon their needs. From each analytical "workspace", a dedicated connection can be created to the backend GROUSE database where researchers have full visibility to multiple schemas and can choose to either query from the original CMS schema or a transformed CDM schema.[21]

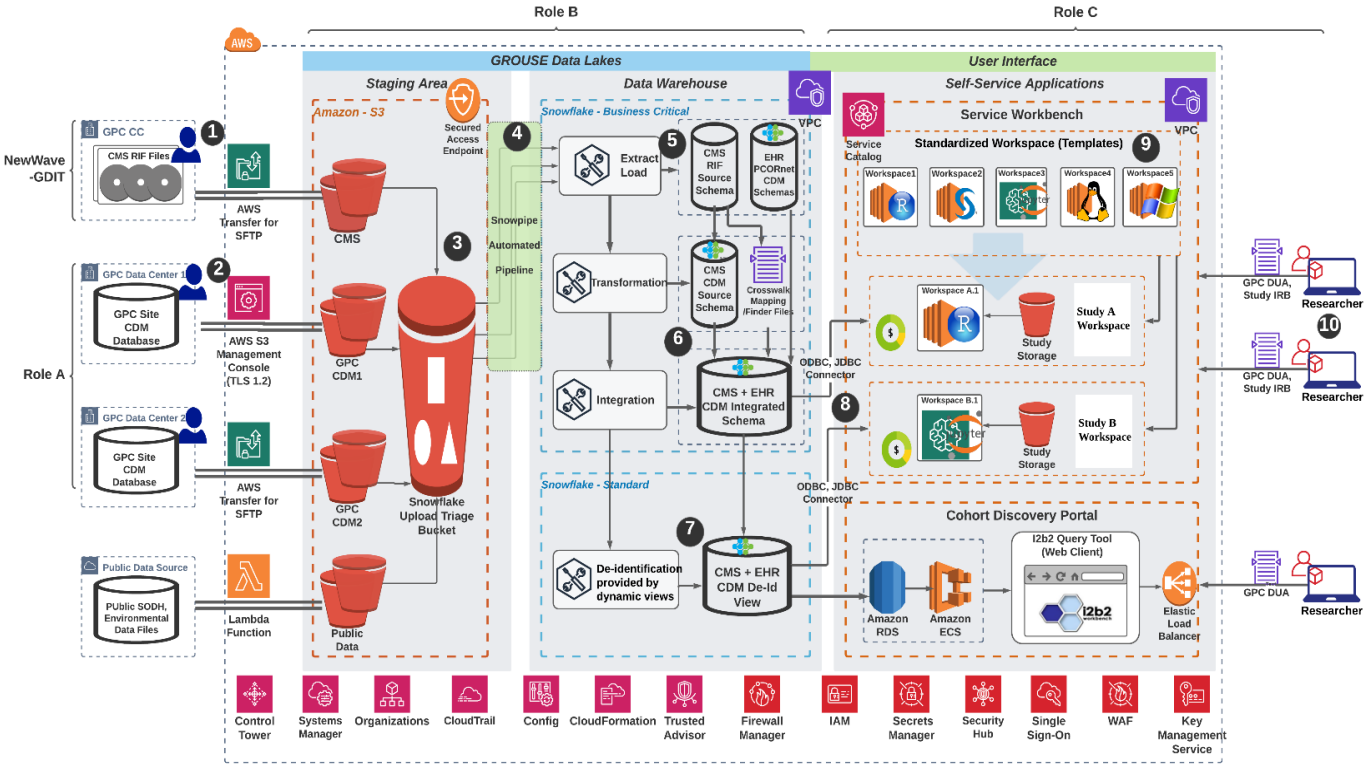

Figure 5*. GPC Cloud-based Data Enclave using Snowflake and AWS*

CMS Data Privacy Safeguard Program Data Management Plan Self-Attestation Questionnaire

Organizations preparing a CMS Data Use Agreement (DUA) application using RIF data must complete a Data Management Plan Self-Attestation Questionnaire (DMP SAQ) to demonstrate compliance with CMS security and privacy requirements. This new procedure piloted by CMS in 2020 provides more structured and consistent guidance on DMP development and addresses cloud computing. The DMP SAQ better matches with the National Institute for Standards and Technology Special Publication (NIST SP 800-53), which dictates the necessary security and privacy controls for federal information systems and provides organizations a process for selecting controls to protect organizational operations and assets. The DMP SAQ is reviewed and renewed annually by the CMS Data Privacy Safeguard Program (DPSP) consisting of 3rd-party auditors from MBL technologies and its subcontractors. The full DMP SAQ and current recertification notice will be provided as supplemental files. More specifically, our DMP SAQ process included the steps:

a) Risk Categorization and Data Classification: we started with classifying data based on Federal Information Processing Standards (FIPS) 199&200 with our existing institutional Data Classification Level (DCL) policy;

b) Control Selection and Implementation: we then identified and adopted required controls specified in DMP SAQ from 18 control families established in NIST SP 800-53 guidelines, as the official policy for the GROUSE infrastructure. These controls are applied to not only the system housing the sensitive data, but also enterprise functions supporting it (e.g., separation of responsibility, control management);
c) System Security Plan: we developed a complete system security plan which entails how the NIST requirements will be met and completed the DMP SAQ;
d) Evidence Gathering and Independent Assessment: once the system is built based on the system security plan, we collected evidence and engaged the CMS DPSP to perform independent assessment. CMS provides an information technology concierge who provides feedback during the development of the DMP SAQ.

Data Confidentiality

Information about study subjects will be kept confidential and managed according to the requirements of the Health Insurance Portability and Accountability Act of 1996 (HIPAA). Data Management Plan following CMS DMP SAQ in compliance with NIST SP 800-53 must be enforced at all time by all approved study team members. Only fully de-identified dataset with approved cohort and data elements (e.g., exclude opt-out sites) will be released for research purpose and accessible to approved and trained study team members.

Waiver of Consent

Describe how the research could not practicably be carried out without the requested waiver or alteration.

With the amount of patients' data needed for these analyses, there is a lack of the time and fiscal resources to obtain patient-level authorization for each individual.

Describe how the waiver or alteration will not adversely affect the rights and welfare of the participants.

No PHI information will be used (in restricted and secure fashion) for any other purposes except for linking to CMS data. Only limited datasets (LDS) are allowed to directly transfer from different data providers (e.g., GPC sites, CMS CCW) to GPC CC in a secured fashion (e.g., encryption-in-transit). Only fully de-identified dataset with approved cohort and data elements (e.g., exclude opt-out sites) will be released for research purpose and accessible to approved and trained study team members. Loss of data confidentiality is considered minimal.

Data Storage and Preservation

Data will be retained for as long as the PCORnet contract and the keystone DUA. It is the responsibility of GPC CC to inform the investigator when these data no longer need to be retained.

Data Analysis and Result Dissemination

The analytic methods and code developed to support PCORnet will be shared as open-source materials on our GPC websites to facilitate adoption and dissemination with the PCORnet and potential investigators who might use the PCORnet resource. Our summarized analytic and quality control results (adhering to the cell size suppression policies) will be shared with the PCORnet coordinating center and PCORI.

Results from the study will be targeted for publications in peer-reviewed journals for each of the cohorts and the informatics methods employed. Prior to publication, the results may be presented at a national or international scientific meeting, such as the annual AMIA Joint Summits on Translational Science, Association for Clinical and Translational Science, San Antonio Breast Cancer Symposium, and the American Association of Neuromuscular & Electrodiagnostic Medicine Annual Meeting, etc.

References

1. Rowland LP, Shneider NA. Amyotrophic lateral sclerosis. N Engl J Med. 2001;344(22):1688-700.

2. Logroscino G, Piccininni M, Marin B, Nichols E, Abd-Allah F, Abdelalim A, et al. Global, regional, and national burden of motor neuron diseases 1990–2016: a systematic analysis for the Global Burden of Disease Study 2016. The Lancet Neurology. 2018;17(12):1083-97.

3. Mehta P, Raymond J, Punjani R, Larson T, Bove F, Kaye W, et al. Prevalence of amyotrophic lateral sclerosis (ALS), United States, 2016. Amyotrophic Lateral Sclerosis and Frontotemporal Degeneration. 2021:1-6.

4. Hardiman O, Al-Chalabi A, Chio A, Corr EM, Logroscino G, Robberecht W, et al. Amyotrophic lateral sclerosis. Nature Reviews Disease Primers. 2017;3(1):17071.

5. Xu X, Shen D, Gao Y, Zhou Q, Ni Y, Meng H, et al. A perspective on therapies for amyotrophic lateral sclerosis: can disease progression be curbed? Translational Neurodegeneration. 2021;10(1).

6. Sun Y, Li X, Bedlack R. An evaluation of the combination of sodium phenylbutyrate and taurursodiol for the treatment of amyotrophic lateral sclerosis. Expert Rev Neurother. 2023:1-7.

7. Miller RG, Jackson CE, Kasarskis EJ, England JD, Forshew D, Johnston W, et al. Practice parameter update: the care of the patient with amyotrophic lateral sclerosis: multidisciplinary care, symptom management, and cognitive/behavioral impairment (an evidence-based review): report of the Quality Standards Subcommittee of the American Academy of Neurology. Neurology. 2009;73(15):1227-33.

8. Group TACtsApI-IS. The Amyotrophic Lateral Sclerosis Functional Rating Scale. Assessment of activities of daily living in patients with amyotrophic lateral sclerosis. The ALS CNTF treatment study (ACTS) phase I-II Study Group. Arch Neurol. 1996;53(2):141-7.

9. Gordon PH, Miller RG, Moore DH. ALSFRS-R. Amyotroph Lateral Scler Other Motor Neuron Disord. 2004;5 Suppl 1:90-3.

10. Simmons Z, Felgoise SH, Bremer BA, Walsh SM, Hufford DJ, Bromberg MB, et al. The ALSSQOL: balancing physical and nonphysical factors in assessing quality of life in ALS. Neurology. 2006;67(9):1659-64.

11. Moore SR, Gresham LS, Bromberg MB, Kasarkis EJ, Smith RA. A self report measure of affective lability. J Neurol Neurosurg Psychiatry. 1997;63(1):89-93.

12. Verma A, Steele J. Botulinum toxin improves sialorrhea and quality of living in bulbar amyotrophic lateral sclerosis. Muscle Nerve. 2006;34(2):235-7.

13. Anagnostou E, Rentzos M, Alexakis T, Zouvelou V, Zambelis T, Evdokimidis I. Volume matters: the influence of different botulinum toxin-A dilutions for sialorrhea in amyotrophic lateral sclerosis. Muscle Nerve. 2013;47(2):276-8.

14. NINDS. NINDS Common Data Elements. Amyotrophic Lateral Sclerosis. [Available from: http://www.commondataelements.ninds.nih.gov/ALS.aspx#tab=Data_Standards.

15. Institute NC. SEER Stat Facts Sheets: Breast Cancer. 2022 [Available from: http://surveillance.cancer.gov/devcan/.

16. Society AC. Cancer Facts & Figures 2023 2023 [Available from: https://www.cancer.org/content/dam/cancer-org/research/cancer-facts-and-statistics/annual-cancer-facts-and-figures/2023/2023-cancer-facts-and-figures.pdf.

17. Polyak K. Heterogeneity in breast cancer. J Clin Invest. 2011;121(10):3786-8.

18. Cawley J, Meyerhoefer C. The medical care costs of obesity: an instrumental variables approach. J Health Econ. 2012;31(1):219-30.

19. Tran BX, Nghiem S, Afoakwah C, Latkin CA, Ha GH, Nguyen TP, et al. Characterizing Obesity Interventions and Treatment for Children and Youths During 1991-2018. Int J Environ Res Public Health. 2019;16(21):4227.

20. Wang LY, Chyen D, Lee S, Lowry R. The association between body mass index in adolescence and obesity in adulthood. J Adolesc Health. 2008;42(5):512-8.

21. Waitman LR, Song X, Walpitage DL, Connolly DC, Patel LP, Liu M, et al. Enhancing PCORnet Clinical Research Network data completeness by integrating multistate insurance claims with electronic health records in a cloud environment aligned with CMS security and privacy requirements. Journal of the American Medical Informatics Association. 2021.