Distribution Agreement

Signature:

Cameron England Date

Case Study

on an Inter-Institutional EHR-Linked HIV Disease Registry

in the Southeastern United States, 2018

By

Cameron England
Master of Public Health
Applied Public Health Informatics

Executive MPH

Laura M. Gaydos, Ph.D.
Committee Chair

J. Mark Conde
Committee Member

Minh L. Nguyen, M.D./M.P.H.
Committee Member

Case Study
on an Inter-Institutional EHR-Linked HIV Disease Registry
in the Southeastern United States, 2018

By

Cameron England

Bachelor of Science
Georgia State University
2004

Thesis Committee Chair: Laura M. Gaydos, Ph.D.

An abstract of
A thesis submitted to the Faculty of the
Rollins School of Public Health of Emory University
In partial fulfillment of the requirements for the degree of
Master of Public Health in the Executive MPH program
2019

ABSTRACT

Case Study
on an Inter-Institutional EHR-Linked HIV Disease Registry
in the Southeastern United States, 2018

By Cameron England

Background. This case study explores an HIV disease registry developed at Emory Center for AIDS Research with a healthcare partner that demonstrates successful inter-institutional sharing of protected health information. Secondary uses of patient data collected in electronic health systems have valuable, broad applications in public health. A common challenge is that healthcare organizations lack the skill, knowledge and resources to leverage this data for secondary uses. Furthermore, a defensive environment exists for sharing HIPAA-protected patient information because of legal and financial consequences. Researchers can help provide the necessary resources; however, negotiating data access is the primary challenge in building a disease registry. This case study demonstrates a pathway for sharing patient data between two institutions by examining the characteristics that influence the organizational behaviors, requirements, goals, and relationships.

Methods. The case study is formulated with a multi-modal approach of a descriptive case study that incorporates iterative stakeholder interviews, protocol analysis, observations, review of documents and archived records, process evaluation, and exploring the physical environment. Inter-institutional data agreements were also reviewed to understand the legal partnership.

Results. The disease registry was developed within the healthcare organization's informatics enterprise, so the data stewards maintain control over patient data. Data are migrated from several data sources that include EHR, LIMS, and pharmacy databases. ETL processes transfer five domains of data that encompass outpatient visits, patient admissions, medications, lab results, and procedures that resulted in nine relational tables contained in the Oracle database. The database constitutes HIV patients seen at the clinic since 2010 as well as historical data on these patients going back to 2000.

Summary. Key characteristics that contributed to a successful sharing of patient information include: (1) Researchers provide knowledge, skills and experience to manage data for secondary applications thus shifting the burden of work from the healthcare system. (2) The disease registry exists within the healthcare enterprise so data stewards maintain control of uses and security. Furthermore, data migration is unidirectional thus limiting strain on and preventing modifications to the health applications. (3) Emory CFAR ensures the quality of data is scientifically robust and quickly accessible. (4) Accountability processes manage and control uses of data with limited involvement from the healthcare system. (5) Governance strategies safeguard data from impropriety. (6) Security for the database is HIPAA-compliant to ease concerns for allowing an external partner to manage data.

Case Study
on an Inter-Institutional EHR-Linked HIV Disease Registry
in the Southeastern United States, 2018

By

Cameron England

Bachelor of Science
Georgia State University
2004

Thesis Committee Chair: Laura M. Gaydos, Ph.D.

A thesis submitted to the Faculty of the
Rollins School of Public Health of Emory University
In partial fulfillment of the requirements for the degree of
Master of Public Health in the Executive MPH Program
2019

Table of Contents

1 DEFINITION OF TERMS

AIDS	Acquired Immunodeficiency Syndrome
ARRA	American Recovery and Reinvestment Act of 2009
BAA	HIPAA Business Associate Agreement
CCO	Chief Compliance Officer
CD4	CD4 lymphocytes test
CDC	Centers for Disease Control and Prevention
CFAR	Emory U. Center for AIDS Research
CITI	Collaborative Institutional Training Initiative
CKD	Chronic Kidney Disease
DUA	Data Use Agreement
EC	Executive Committee
EHR	Electronic Health Records System
ETL	Extraction-Transformation-Loading Processes
FTP	File Transfer Protocol
HIV	Human Immunodeficiency Virus
HIPAA	Health Insurance Portability and Accountability Act of 1996
HITECH	Health Information Technology for Economic and Clinical Health Act
HRSA	U.S. Health Resources and Services Administration
IRB	Institutional Review Board
IT	Information Technology
IRB	Institutional Review Board
LIMS	Laboratory Information Management System

LITS	Emory University Library and Information Technology Services
LOINC	Logical Observation Identifiers Names and Codes
MOU	Memorandum of Understanding
MRN	Medical Record Number
NIH	National Institutes of Health
OLAP	Online Analytical Processing
OLTP	Online Transactional Processing
OMOP	Observational Medical Outcomes Partnership
OTT	Emory U. Office of Technology Transfer
PHI	HIPAA-Protected [Patient] Health Information
PLWH	People Living With HIV
PLWHA	People Living With HIV and AIDS
PMP	Project Management Plan
RAC	Research Advisory Council
RHS IT	Emory U. Research & Health Sciences Information Technology
RSPH	Emory U. Rollins School of Public Health
SME	Subject Matter Expert
SNOMED	Systematized Nomenclature of Medicine – Clinical Terms
SOM	Emory U. School of Medicine
SQL	Structured Query Language
VL	Viral Load Test
VPN	Virtual Private Network

Case Study on an Inter-Institutional EHR-Linked HIV Disease Registry in the Southeastern United States, 2018

2 Executive Summary

HIV infection is a high value target for public health because of its high long-term cost of care and its disproportionate impact on men that have sex with men, Black/African American people, and the poor. Outcomes-related research from diverse health systems is useful to collect information about people living with HIV/AIDS to control the spread and improve interventions. This case study explores an HIV disease registry developed at Emory Center for AIDS Research (CFAR) with a Southeastern United States healthcare partner in a relationship that demonstrates successful inter-institutional sharing of protected health information. Researchers and healthcare leaders are increasingly finding value in secondary uses of patient health information using protected patient care data that are collected in electronic health systems.

Navigating the requirements to support inter-institutional data exchange is complex. A common challenge is that healthcare organizations lack the skill, knowledge and resources to leverage the use of patient health data for public health. Furthermore, healthcare organizations tend to take a defensive position in sharing HIPAA-protected patient information because of legal and financial consequences resulting from security breaches of protected health information. Researchers contribute the skill and knowledge, but gaining access to data is a common barrier because they are not employed by the healthcare organizations from whom they wish to utilize data for public health research. Researchers are often funded by grants that can help provide the resources to create a disease registry, thus shifting the work, skills and resources of data discovery from the healthcare organization. However, researchers often discover that negotiating data access is the more significant challenge in building a disease registry. This case study demonstrates a pathway for developing an inter-institutional relationship for sharing patient data by preparing strategies for interoperability, governance, accountability, and security to present a transparent process that assuages the concerns of the healthcare organization. The study aims to describe a complex data sharing case in its real-life, natural context to gain broader understanding of the complexities of sharing protected health information across organizations.

The methodology follows a descriptive case study format with which to evaluate processes, stakeholders, physical framework, security, and governance of a disease registry linked to electronic health records. Inter-institutional data agreements were also reviewed to understand the legal partnership. This study presents a method to evaluate the characteristics of the disease registry and the group that developed and manages the disease registry as well as understand how the organizational behaviors, requirements, goals, and relationships are aligned to support an inter-institutional relationship for sharing HIPAA-protected data. The case study is formulated with a multi-modal approach incorporating iterative stakeholder interviews, protocol analysis, observations, review of documents and archived records, and exploring the physical environment.

The HIV disease registry was developed and implemented using a $200K grant from the NIH. Emory CFAR has committed to cover the annual operating costs of approximately $150K. As of July 2018, the data in the registry included 13,033 unique patients since 2010 with 540,143 encounters representing 80,775 human years. There are 348,731 CD4 and viral load lab results that researchers use as key markers for HIV disease progression. Currently, 176 covariates exist in the registry divided across nine relational data tables: diagnoses, encounters, lab results, dispensed medications, ordered medications, medical record number, person, problem lists, and intake procedures. The registry functions under an IRB-approved protocol and a data management agreement with the healthcare organization. The disease registry is governed by an executive committee and research advisory council involved in approving all uses of patient data.

The HIV disease registry is contained within the healthcare organization's informatics enterprise, allowing the data stewards to maintain a majority of control over patient data. Data are migrated from several data sources that include the electronic health records, laboratory information management systems, and pharmacy database. Only patients with diagnosed HIV disease are contained within the registry, but these were not limited to only patients presenting to the HIV clinic. The patient data that are collected were expanded to include any HIV-diagnosed patient throughout the healthcare organization. Five domains of data are migrated from the healthcare data sources that encompass outpatient visits, patient admissions, medications, lab results, and procedures. The data from these domains are then parsed to the nine relational tables organized within the Oracle-based data system. The patient population contained within the registry constitutes HIV patients seen at the healthcare system since 2010 but also incorporates historical data on these patients going back to 2000.

In depth exploration of the HIV disease registry to discerned key characteristics that contributed to a successful inter-institutional sharing of HIPAA-protected patient information. These include:

(1) Emory CFAR provides the knowledge, skills and experience to manage patient data for secondary applications thus shifting the burden of work and resources from the healthcare system.

(2) The disease registry was installed within the healthcare system informatics enterprise so the data stewards would maintain control of data uses and security. Furthermore, data migration is unidirectional from the healthcare applications to the disease registry thus limiting strain on and preventing modifications to the systems used for managing care.

(3) Emory CFAR affords continuity to ensure the quality of data is scientifically robust and provide investigators with quick access to healthcare data for secondary analyses. Continuity is gained because Emory CFAR is a research organization with long-term funding from the NIH, whereas investigators typically would have National Institutes of Health (NIH) research grants for less than 5 years.

(4) Accountability processes were implemented for the disease registry to manage and control uses of patient health information without deep involvement from the healthcare system alleviating stress on resources intended for care rather than research.

(5) Scientific and technical governance strategies enacted will safeguard data from impropriety and facilitate publishable research.

(6) Security for the database and data are compliant with physical, administrative and technical security requirements of HIPAA to ease concerns of the data stewards for allowing an external partner to manage data owned by the healthcare organization.

Organizations and researchers can use the Emory CFAR HIV Disease Registry as a model for building new disease registries. The versatile framework and schema make this a good, low-cost solution with for constructing a disease registry for other diseases. The HIV disease registry has proven to be a successful model in developing new disease registries for reproductive health, digestive diseases, and transgender health. The key strategies outlined in the case study have forged stakeholder acceptability for the HIV disease registry allowing for successful inter-institutional sharing of HIPAA-protected patient health information to facilitate secondary uses of patient health to support public health, clinical and translational research, programmatic evaluation, and health outcomes assessment.

3 INTRODUCTION

3.1 BACKGROUND

This case study evaluates an HIV disease registry to demonstrate a successful inter-institutional relationship in which protected health information can be shared. This is particularly of interest and useful for academic research groups that work within a non-academic healthcare environment. This case addresses governance, accountability and security processes that have allowed the HIV disease registry to overcome a defensive environment to explore judicious secondary uses of patient health data.

Widespread adoption of electronic health records systems (EHR) has stimulated more demand for secondary uses of patient health data. The appeal of newly accessible electronic health data has driven investment for solutions to overcome new challenges of translating vast complex medical data into meaningful information. Electronic health data can be leveraged for secondary uses that will contribute to organizational objectives for patient care management, public health, clinic operations, and research. With healthcare policy shifting to performance-based payments targeting improvements to patient health outcomes, it will be even more critical for healthcare organizations to use patient health data to effectively and quickly track, assess, and report successes in delivery of care. Because healthcare accumulates massive amounts of clinical transactional data, EHRs represent an important source of information that can be beneficial to broad applications in healthcare and public health initiatives. However, to capitalize on the benefits of using patient health data, researchers will have to overcome common challenges to sharing data in the healthcare industry to be able to maximize the utilization of protected health information while ensuring patient confidentiality and propriety.

Disease registries offer an informatics solution to support secondary uses of patient health data. The disease registry is a specialized database containing information about a targeted population of individuals diagnosed with similar conditions or diseases. These registries can provide more effective architecture than EHRs for secondary uses in clinical informatics. The registry framework is more adept at connecting disparate data sources and heterogeneous data using an organized approach that can exploit the flexibility and versatility of its data infrastructure to transform valuable health data into actionable information. Disease registries, like data warehouses, extract and manage copies of health data using separate hardware and applications that are configured for research-driven efficiency and performance that affords security, minimal impact on the EHR and healthcare processes, diminished strain on resources needed for EHR data services, and enhanced query performance. Data are transformed, loaded and aggregated into meaningful tables in a relational database where it can be more easily managed, manipulated and queried.

Government agencies, healthcare organizations, and their partners can improve disease management through secondary analysis of EHR-derived patient data to enhance surveillance, research, prevention, and evaluation activities [1]. A disease registry supplementing the new EHR system will offer insight into the capability and usefulness of transactional health data that can meet demands for big data management, meaningful use requirements, as well as national care guidelines for diseases and conditions. This type of system is capable of rapid access and mining of data to answer research questions and evaluate quality metrics. EHRs cannot perform data searches as efficiently and their frameworks struggle with longitudinal data extractions. In addition to quality assessments, disease registries have successfully been used for a variety of real-world applications that include feasibility determination, strengthening grant applications, case identifications, aggregate patient-disease-population level statistics, programmatic evaluations, data exploration and discovery that will lead to improvements in the delivery of care and public health.

A Southeastern U.S. academic research facility has an enduring partnership with a local healthcare organization to provide medical education, patient care, and research. One of the largest public hospitals in the Southeast, this healthcare organization is a community safety net hospital with an historic commitment to the health needs of underserved populations in the region. The hospital maintains a Ryan White-funded clinic for specialized HIV ambulatory care in metro-Atlanta that serves the sickest population of people living with HIV (PLWH) and AIDS (PLWHA). The Ryan White Care Act of 1990 provides payer of last resort financial support for HIV care and treatment to the poor, indigent, uninsured and underinsured. This clinic is an internationally recognized and respected patient-centered model for comprehensive care and treatment and is among the largest facilities in the nation dedicated to caring for people with HIV/AIDS.

Providing care to over 6,000 patients, the HIV clinic is responsible for a high share of poor minorities with HIV/AIDS from residents of 22 Georgia counties that represent one in five PLWHA in Georgia and a third of all PLWHA in metro Atlanta [2]. Almost all patients at the HIV clinic live below 200% of the Federal Poverty Level, but the vast majority (76%) live below 100% of the Federal Poverty Level [2]. Its 78% minority population is nearly exclusively patients with an AIDS diagnosis. This group contends with treatment compliance issues influenced by ecosocial determinants of health involving housing instability, food insecurity, drug addiction, non-supportive social and work environments, incarceration, insurability bias, as well as a high degree of social and class stigma for being poor with HIV. The data collected for the disease management of a large urban AIDS-stricken population are an important tool in understanding the HIV epidemic at local, regional and national levels and for providing important insights into the impact of U.S. HIV care policies on the health outcomes of its target population.

Healthcare organizations can add value to its strategies by using a data registry or warehouse to connect information from heterogeneous healthcare data sources. Patient-centered EHRs are designed to support care and billing transactions during clinical encounters and are often ill adapted for secondary uses. Healthcare organizations need successful integration and harmonization of internal and external data sources in order to meet the 2009 American Recovery and Reinvestment Act (ARRA) requirements for meaningful uses and to allow full realization of the value of patient health information in supporting the organizational mission. EHRs mostly exist in proprietary platforms that lack the capacity for broad interoperability and scalability to connect other disparate data sources that exist within the same healthcare enterprise that include EHRs, laboratory information management systems (LIMS), immunization records, clinical research systems, and pharmacy systems.

Cross-organizational access of HIPAA-protected patient health information is complex beyond the collection and aggregation of data from disparate sources. Stakeholder buy-in and identifying high-level champions at the healthcare organization were among the most significant challenges experienced by the HIV disease registry project team. Over the course of four years, Emory Center for AIDS Research (CFAR) has successfully developed a relationship to access and use patient health data owned by the healthcare organization. The CFAR at Emory University is one of 19 centers in the United States that are funded by the National Institute of Allergy and Infectious Diseases to support multidisciplinary research at academic institutions aimed at reducing the burden of HIV both in the Unites States and around the world. However, there remains a fractured relationship with the regional healthcare organization that constrains the use of patient health data to support the Emory CFAR mission. The challenges expressed by Emory CFAR's public health investigators for secondary uses of patient health data are listed here. These issues are further addressed in the results with descriptions for how the challenges were overcome by Emory CFAR.

- **Technical Capability:** The healthcare organization does not have the experience and ability to support research uses of data. In addition, EHRs continue to struggle with using data beyond patient care and billing. The EHR used by the healthcare organization is valuable for cross-sectional analysis, but in most cases, longitudinal evaluations are not feasible. In comparison, the HIV disease registry is designed to collect discrete and flexible retrospective, longitudinal, and cross-sectional views of case- and population- levels of clinical data.
- **Data Access:** Investigators do not have quick or comprehensive access to data. In addition, the expertise divide between researchers and information technology (IT) are a hindrance. IT staff do not fully understand clinical characteristics of diseases and researchers do not fully understand what data are available so therefore find it difficult to develop an efficient data collection strategy. Another issue that has arisen is modifications or changes to lab or diagnosis codes. It is rare that these changes are communicated to researchers and at times is not communicated between hospital departments so the data that is requested may not be sufficiently comprehensive, representative or up to date. Another issue is that the healthcare organization did not have a defined pathway to share data with a disease registry or external partners.
- **Institutional Strategies for Patient Information:** The two institutions, Emory CFAR and the healthcare organization, have different mission goals for the use of patient health data. While Emory CFAR supports clinical and public health research, the healthcare organization is focused on patient care, quality initiatives and programmatic evaluations that can provide evidence-based support for clinical care. Dedicated hospital resources for research functions are greatly limited. The healthcare organization has a rationally cautious and defensive environment for sharing data and relinquishing control with external partners.

- **Competition for Resources:** Related to the differing institutional strategies, many Emory research groups vie for the few resources to collect data. The healthcare organization's IT and business intelligence teams are prioritized for hospital operations over research data requests, so Emory research groups have to compete with each other to obtain data and often may have long and unpredictable wait times to obtain data as the healthcare organization is focused on optimizing care operations and delivery. In the world of NIH application and study deadlines, long waits and unpredictability for data access introduces a significant barrier to Emory investigators needing the information to meet NIH obligations.

3.2 Problem Statement

An academic research center is engaged with a Southeastern U.S. infectious disease clinic to leverage patient health data for secondary uses that support clinical and public health research, programmatic evaluations and process improvements that will promote better health outcomes for HIV patients receiving care at the clinic. There is a need to address issues in gaining access to protected health information owned by an external healthcare organization, securely organizing data to facilitate secondary uses of patient health data, and providing resources and programs to support public health activities that present researchers with access to patient health data. This case study examines the Emory CFAR HIV disease registry's establishment of a process that shifts the work of data discovery and management away from the healthcare organization to experts in HIV disease research.

3.3 Purpose Statement and Public Health Impact

This case study evaluates a versatile informatics platform that supplements the EHR as a viable business and research tool capable of rapidly addressing key needs of investigators and healthcare leaders to support their respective missions. The principal objective of this case study is to review an HIV disease registry to assess how a cross-organizational data sharing relationship was contemplated and successfully implemented. Using this case study, recommendations can be derived for other clinical-academic partnerships seeking to address similar issues in comparable technology environments.

4 LITERATURE REVIEW

Healthcare data assets have become valuable tools in public health research. Beyond the obvious uses of patient health data for direct clinical care, these data may also be used for purposes other than patient care, such as disease surveillance, identification of at-risk populations, determining treatment effectiveness, quality evaluation, health management, cost analysis, and programmatic evaluation [3]. According to a 2009 Price Waterhouse Coopers survey, the majority of healthcare respondents say they support some form of secondary data use of patient health data while 90% of industry respondents believe secondary uses of these data have the potential to significantly improve patient care and offer even greater benefits in the future. The report suggests an approach to meaningful use of health data should target sharing the minimal amount of data as possible and in-demand data on specific disease conditions [4]. Disease registries would offer a solution such as this.

Government agencies, healthcare organizations, and their partners can manage diseases, such as HIV/AIDS, by employing EHR patient data to enhance surveillance, research, prevention, and evaluation activities [1]. By the end of 2015, the Centers for Disease Control and Prevention (CDC) estimated that 1.1 million people are living with HIV in the U.S, and 15% are unaware of their infection [5, 6]. HIV/AIDS is a high value target for change because it is a significant driver of burden on quality of life and high lifelong health and treatment costs [7-9]. Measuring the effectiveness of HIV treatment has become a prominent metric for success in limiting the spread of HIV infection and improving health of HIV-infected individuals [6, 10, 11]. The 2013 HIV Care Continuum Initiative was launched as a supplement to improve and quantify the goals set forth by the 2010 National HIV/AIDS Strategy of the U.S. White House. Since its implementation, research has revealed significant gaps in HIV care despite the efforts of a mobilized, well-funded nation. Even with the invested resources of the U.S. government for HIV prevention and care, only 25% of people living with HIV (PLWH) in the U.S. have controlled the disease [12]. As a result, this national strategy sought to elucidate the underlying issues for the staled progress in the nation's battle against the HIV/AIDS epidemic. New case definitions for the care cascade aided in inaugurating key metrics for determining success at the critical stages of HIV care: linkage to care, retention in care, prescription of HIV treatment, and viral suppression. These metrics along with the changing healthcare legislation landscape are guiding the needs for a specialized HIV disease registry for CFAR.

In response to the 2010 National HIV/AIDS Strategy, the Fulton County Department of Health in Georgia has issued its own localized strategy to support HIV population health [13]. With a growing and aging population of PLWH in metro Atlanta, efforts have been made to improve the quality of care delivered and to more effectively manage health care resources in metro Atlanta. These initiatives rely on being able to rapidly access health data to accurately measure and analyze performance in care delivery. According to the Georgia Department of Public Health Atlanta area researchers and healthcare leaders have already been leveraging common data sources at healthcare enterprises for secondary uses of patient health data [14].

Being able to develop successful public health initiatives relies on the quality of data to formulate evidence-based knowledge. Because healthcare accumulates enormous amounts of data from clinical encounters, EHRs represent an important tool to leverage information for the Big Data initiatives in healthcare that can further advance knowledge in medicine and public health leading to improved health outcomes through evidence-based decision support [15]. Furthermore, health data offer an opportunity to partner key users of information in a way that cultivates new cross-disciplinary collaborations with clinicians, scientists, consumers, pharmacies, payers, employers, and technology providers all working together in tackling complex research [16]. With EHRs at 87% percent adoption among US office-based physicians [17], efficiencies from healthcare informatics technology has the potential to offer $77 billion in annual savings from patient care by reducing lengths of hospital stays, nursing time, and drug usage in outpatient and inpatient care [18].

Leading the incitement for growing EHR adoption have been the 'meaningful uses' that are enforceable mandates by the Health Information Technology for Economic and Clinical Health Act (HITECH) supplement of the 2009 ARRA legislation. This legislation has dispensed over $27 billion in financial incentives to healthcare organizations to promote EHR implementations for effective uses of patient health data. The move to EHRs presents significant opportunities in public health, but requires a high level of integrity and security to manage processes and system capabilities for patient data research. Included in HITECH is new provisions to strengthen HIPAA rules to ensure privacy and security in data management and transfers. These rules levy high penalties for data misuse or negligence, so many data owners have defensive attitudes around the sharing of protected health data. The security, regulations and criminalization of data mishandling are important obstacles that not only governs how healthcare enterprises are built, but is also ranked among the highest and costliest healthcare concerns that impedes interoperability and secondary uses of health data [18].

Public health research is dependent on data that are accurate, comprehensive, timely, and accessible. It is inherently difficult to improvise a solution that balances all these needs and still delivers robust scientific-quality data. Many obstacles exist that include the proprietary nature of EHR platforms that restrict the ability to access and use transactional health data for secondary uses, the potential to compromise the efficiency of systems that deliver direct care functions, and the relational data models of EHRs are not structured to effectively support secondary uses [19, 20]. The challenges are further complicated by lack of standards, organizational capacity, privacy concerns, and technological limitations [21]. Huge volumes of data raise concerns about quality as well as consistency, standardization and a controlled vocabulary for sharing data [1]. Several recurring issues for secondary uses of health data have been the completeness and accuracy of data stored in clinical databases [22-25], the effective capture and assessment of unstructured contextual data that exists in EHRs [26-28], the diverse heterogeneity of internal and external data sources, and ensuring data values are meaningful and consistent [29]. In many cases the researcher has little or no influence on how data are collected, entered or maintained [30].

Furthermore, EHR data do not automatically translate to knowledge, but rather have to be manipulated and transformed into a useful form with meaningful organization that better facilitates secondary uses. Often data requirements for EHRs are not ideally framed for secondary uses of electronic patient health information leading to poor data quality for non-primary uses [31]. The scope that is intended for basic EHR architecture typically lacks the amplitude needed to carry the full responsibilities of an informatics enterprise capable of fully leveraging secondary use opportunities. Many systems struggle to successfully integrate and harmonize data from internal and external heterogeneous data sources in ways that can meet ARRA's meaningful use requirements. They lack the capacity for broad interoperability and scalability to connect disparate data sources that can exist within the same healthcare enterprise. The lack of interoperability diminishes comprehensiveness, validity, speed, and accuracy of available data [32-34]. These are significant challenges for any healthcare organization and requires specific expertise and knowledge that were not included in the clinical care infrastructure before the adoption of electronic data.

A 2010 evaluation of clinical registries, administrative databases and EHRs reported that when used properly sources with large amounts of data were useful in improving health outcomes for patients with colorectal cancer [25]. However, Logan and Lieberman described the limitations of data that can often be inaccurate and incomplete. Their study results demonstrate the two methods of data collection, automated and manual, were shown to suffer from different biases. Automatically extracted data result in errors that are more systematic. Inaccuracies found in manually extracted data were found to lean towards random errors. Another key limitation was that much of the data captured in EHRs were narrative data that are difficult to use for targeting and capturing discrete information. The study summarized that successful use of these data sources for improving quality of care required the investigator to have a complete understanding of the data being collected and how it is collected to fully comprehend the limitations of accuracy and completeness.

Disease registries offer a viable informatics solution that best supports secondary uses. Registries exist as a specialized relational database targeting health information collected about a patient population with a specific disease or condition. The EHR takes a one-patient-one-record approach with real-time on-line transactional processing (OLTP) designed to efficiently support care delivery and billing transactions. This works well for managing patient care because it is modeled from a clinical encounter-centric perspective with a data structure to support financial transactions and analyses, however, is poorly designed for research functionality that supports secondary uses. The disease registry is more effective for secondary uses because it operates from a versatile object-oriented perspective based on on-line analytical processing (OLAP) architecture that is more adept at managing heterogeneous electronic data from disparate sources to better facilitate secondary uses in clinical informatics [35].

OLAP can be exploited for its flexible and versatile data infrastructure to transform valuable health data into actionable information. Disease registries, similar to data warehouses, extract and manage copies of health data using hardware and applications that are separate from the EHR so that it affords better security with minimal impact on the EHR itself and therefore does not affect clinical care operations. There is also diminished strain on resources used for EHR data services when data are separated from clinical care systems to be used for secondary research. In a disease registry, data are transformed, loaded and aggregated into meaningful tables in a relational database where it can be more easily managed, manipulated, and queried. In secondary use the data are often de-identified, anonymized, or translated into a limited data set before it is used or aggregated with other data [36]. Opportunities for secondary uses have led to the development and utilization of disease registries linked to EHRs that are used to supplement patient health care [37].

In 2001, the Cleveland Clinic adopted an EHR for its 9 hospitals and 13 community clinics. Navaneethan et al. were able to leverage patient health data from the EHR to develop a disease registry for chronic kidney disease (CKD) to identify cases, follow the care of these patients to observe outcomes, and to establish a comprehensive resource for interventions related to CKD [38]. This registry was linked to the EHR so that it was capable of capturing patient level data targeting key elements that included demographic, clinical and laboratory data. In 2010, the registry contained over 57,000 patients. The CKD registry was validated in this study by comparing its data with the EHR records to ensure the registry could reliably capture data. A key challenge for the Cleveland Clinic was demonstrating the usefulness of the registry data compared to data that are captured in a clinical study, which creates a more controlled environment. Data about a larger population in an EHR are able to overcome the limitations of clinical studies because clinical study participants are not always representative of the real population. Additionally, EHRs include comprehensive data about a population, such as socioeconomic and insurance data that are typically excluded from a treatment clinical trial. The study team established that an EHR-based disease registry with reliable data was possible in a large academic health system and were able to use this to support research studies, improve the quality of care for CKD patients and begin a CKD surveillance program at Cleveland Clinic.

In Chicago, three public hospitals developed a cross-organizational quality improvement project that established the eID clinical data warehouse to control hospital infection [39]. A single system to manage laboratory, pharmacy, and administrative data did not exist; therefore, a client-server framework was developed based on a Microsoft SQL platform. The project developed a system to automate data extraction, harmonize data from thirteen data sources and connect the information from each hospital on a single, web-accessible server. Manual validation was accomplished by analyzing data on a sample population to determine completeness, continuity and accuracy. This system exhibited similar barriers and challenges as expected when developing a cross-organizational agreement for the management of protected health data. The study team encountered regulatory and political barriers that required negotiating executive-level endorsements from each hospital. They also found that the hospitals lacked technical expertise that restricted the study team's ability to extract the necessary data. In establishing connections to client servers new relationships had to be created with clinical departments to ensure adequate management of data by these groups. Another barrier was lack of documentation for understanding the database model and data dictionary of each of the data sources. Clinical subject matter experts (SME) were consulted to ensure the capture of targeted data elements and sometimes required involvement from the software vendors. Usability of data, or the consistency of how data values are expressed, also created unique challenges, especially when integrating the same data from different sources. In addition, the evolution of the data sources was not typically communicated to the study team, so they were not informed of changes to data or when systems were replaced, modified or upgraded. All of these challenges were similar to what was encountered in the development of the CFAR HIV disease registry.

This case study evaluates the development of the CFAR HIV Disease Registry. The literature review assessed the utility of electronic patient health data used for public health and demonstrates its important role in the progression of U.S. healthcare informatics systems. Challenges and barriers exist in the development of tools for secondary uses of patient health data. Understanding these limitations and lessons learned from other similar systems help to compare and understand the development of the CFAR HIV Disease Registry and the challenges that were faced in its implementation. But despite the challenges, there exist many disease registries linked to EHRs that are currently being used to supplement patient health care indicating there are successful utilization of EHR data [37].

5 METHODOLOGY

5.1 INTRODUCTION

The methodology of this research follows a descriptive case study format with which to evaluate processes, stakeholders, physical framework, security, and governance of a disease registry linked to electronic health records. This project describes a complex data sharing experience in its real-life, natural context to gain broader understanding of the complexities of sharing protected health information across organizations. The environment involves the cross-integration of an academic health center with a public safety net healthcare system to collect disease information about patients living with HIV receiving care at the Southeastern U.S. healthcare system. This study provides a method to evaluate the characteristics of the disease registry and the group that developed and manages the disease registry as well as understanding the organizational behaviors and relationships. Understanding these characteristics and processes can aid other clinical-academic investigators in developing disease registries for other diseases or chronic conditions.

The aims of the methodology were to identify stakeholders and their business requirements, outline the workflows and processes, and review the attributes of the data and the environment. This evaluation incorporated the development and implementation phases of the HIV disease registry. The case study was framed by ongoing stakeholder feedback and investigation to gain a better understanding of the fragmented data sources, the resulting data made available for clinical and public health applications, and also to understand the requirements of key stakeholders and their experiences in developing and managing patient health data using a disease registry in this environment.

This project worked with the managing sponsor, Emory CFAR Clinical Research Core, to gain access to the stakeholders, the database team, and the sources of data used to develop this case study. A review of the literature was also completed to evaluate the role a disease registry has on secondary uses of health data as well as assessing common limitations. The stakeholders' interviews also included identification of constraints related to design, process, and funding. Using this methodology provided a comprehensive understanding of the purpose of this HIV disease registry, its organizational impacts and the strategies needed to enact data sharing.

This case study used a multi-modal approach incorporating interviews, protocol analysis, stakeholder observations, review of documents and archived records, and exploration of the physical environment. The methodology is framed with the feedback of Emory CFAR HIV disease registry's project team and their interactions with key stakeholders. The stakeholder interviews focused on understanding governance, accountability and security for secondary uses of patient health data. Research design also involved review of the business and technical requirements to gain insight into the aspects of data sharing and system interoperability. A review of the physical data model and database framework contributed knowledge about how data are managed. The case study offers a resource for ideas and opportunities for innovation by studying a unique informatics framework that can be translated into knowledge that is more generalizable. This study is not intended to draw a definitive cause-effect conclusion. In addition, these types of study have inherent investigator biases, particularly in this case where the case study investigator was also closely involved in the development, implementation and management of the disease registry.

In general, case studies focus on asking how a phenomenon has occurred. Investigators require access to scientifically robust patient health data that can support public health research that results in evidence-based knowledge to improve health outcomes. For this purpose, investigators are interested in how Emory CFAR was able to successfully produce an informatics solution containing patient health data that it does not own. The case study is elucidating a common problem and informatics need for cross-organizational data sharing by exploring and describing a solution that is in-action.

Illustrative case study research infers an in-depth, detailed assessment of a valuable tool described in such a way that others can become familiar and reproduce the results in their own settings to benefit their initiatives in public health.

5.2 Research Design

Interviews were conducted for this case study to identify key stakeholders, understand their functional and technical requirements for the HIV disease registry, and examine their experiences in a cross-organizational arrangement for sharing patient data. A list of interview questions is included in Appendix I: Interview Questions. The key stakeholders in this case study represented various aspects of the HIV disease registry. Interviews were conducted with the disease registry's principal investigator, informatics architect, data analyst, and clinical subject matter expert (SME). The principal investigator provided insight into governance and business requirements, the informatics architect and data analyst elucidated the framework and technical requirements, while the clinical SME described the data and research strategies. Attendance at the registry meetings also provided valuable insight and uncovered many of the challenges and limitations experienced by the disease registry team.

The HIV disease registry documents were reviewed to assess operational management, governance, security, and business requirements. This case study inspected the business plans, formal evaluations, standard operating procedures, regulatory documents, grant application, and project finances. Among the management documents were included a formal evaluation of business processes conducted by Emory Library and Information Technology Services (LITS) in 2016, the 2013 business case and the 2013 project management plan. Additionally, the case study reviewed the project's budgets from 2013 to 2018. Evaluating these documents explained how data are accessed and managed, the data requirements for the HIV disease registry, how the registry and data sharing are managed, as well as the security controls that are in place.

The case study utilized the business case, the project management plan, standard operating procedures, and the data dictionary to review the technical environment. This allowed the case study to observe the flow and assembly of data from the data sources to the HIV disease registry. To review the physical environment, the case study visited the locations where the hardware were stored. This included the offices that contained the client computer terminals used to access the HIV disease registry. An attempt was also made to visit the physical environment of the network server but was unsuccessful due to physical controls for the healthcare enterprise's technology infrastructure. The cloud storage is a virtual environment that was accessed for this case study by using the client computer terminals.

Evaluation of the data and the system were accomplished through interviews with key stakeholders and by gaining access to the data and database. To access the data content and relational tables this case study used Toad for Oracle v12.6 to run structured query language (SQL) programs designed by this case study investigator. The Toad application was connected to the disease registry using a client computer housed within the healthcare system's physical infrastructure and using Oracle Instant Client v11.2.04 to establish a secured connection between the client computer and the HIV disease registry. SQL programming is necessary to view data in the disease registry, which allowed the case study to search the data in the database to assess its structure and quality. The case study also observed the flow of data between data sources, including where data are stored and where data are in transit.

Among the instruments for this case study was a physical data model that was created by this case study to explore how data are organized within the disease registry. The data model is included in Appendix II: Physical Data Model. Using SQL scripts established the foundation for creating a data model and data dictionary. These scripts allowed the case study to see what data are stored and how they are organized across different tables. A logical data model provided visualization of the relational data tables and the database schema. The case study used Toad Data Modeler 5.3 to design the logical data model. Data modeling is a useful tool to assess the efficiency of how data are organized and normalized within relational tables. This allows the case study to review the quality of data, complexity of the data organization, its normalization structure, and the database's performance efficiency. Additionally, a data model can also provide some insight into the costs associated with managing the data as well as help in understanding the true scope of the disease registry.

5.3 INSTRUMENTS

Table 1: Instruments used to conduct the case study.

Interviews of Key Stakeholders	Principal Investigator	**Governance, Business Requirements, and Research Use:** Described how the registry is managed and its oversight. Described the high-level business requirements and functions of the registry for each of the sponsors. Described opportunities for secondary uses of patient health data afforded by the registry and shared past and current uses of data for clinical and public health applications. Described ongoing challenges in inter-organizational management of patient health data.
	Informatics Architect	**Database Design, Framework and Strategy:** Described the design and development of the database and framework, the technical/informatics relationships with sponsors, the general schema (how data are organized), and the future strategy for expansion and further development of the registry.
	Data Analyst	**Database Design, Framework and Technical Requirements:** Described the logical and physical models of data tables that exist in the database and explained the organization of relational tables and the relationships that exist between the tables. Described the administration and security of the database.
	Clinical Consultant	**Data and Research Strategies:** Described the data strategies used in the development of the registry as well as research strategies for how data are used in clinical and public health applications. Described ongoing challenges specifically related to data content and data quality.

Review Physical & Electronic Documents	Business Case & Requirements	The business case described the concept of the HIV disease registry and its requirements and impacts. This document was used to describe the utility of the registry to organizational stakeholders to provide information with which to make business decisions for the development and implementation of the registry.
	Data Use Agreements	The data use agreements exist between the sponsors to outline how patient health data are managed and can be used for clinical and public health applications.
	Formal Evaluation of Business Processes	The evaluation was conducted by Emory LITS to describe the business processes that include the flow of data, locations of where data exist, the process for accessing data, and the process for how data are shared for clinical and public health applications.
	Project Management Plan	The PMP outlined the development and implementation of the disease registry. This document included the scope management, schedule management, cost management, project organization, communications management, quality management, data governance and security management, risk management, issue management, and procurement management. The PMP was used by the project development team and approved by the sponsors.
	Standard Operating Procedures	SOPs described the various standard processes used in the management of the data in accordance with data use agreements. These described the roles and responsibilities for data and informatics management, essential forms, uses of patient health data, security, risk mitigation, de-identification of data, and requesting data. Also reviewed organizational SOPs for HIPAA Safety and Security Rules describing the use and sharing of patient health data as well as the administrative, technical and physical safeguards.
	Data Dictionary	The data dictionary describes the contents, format and physical structure of data as it exists in the database as well as describes the relationships between various elements. This document is shared with investigators to illustrate the data that are available in the HIV disease registry for secondary research use.
	Regulatory Documents/ Protocols	The regulatory documents include the Institutional Review Board (IRB)-approved scientific protocol, IRB approval letter, and IRB waiver of HIPAA authorization. These documents are used for the governance of the HIV disease registry.
	Grant Application	This document was the application for the petitioning of funds from NIH for the development and implementation of the disease registry in August 2013.

	Budgets & Projections	These documents describe the operational costs of the HIV disease registry.
	Research Data Requests	These were formal requests by investigators to access data from the HIV disease registry for secondary uses of patient health data. The Research Advisory Council approved these requests for data.
	QA Results	These documents included results from the ongoing quality evaluation of data from sources, data contained in the HIV disease registry, and data that are shared with investigators. These were key documents in describing the challenges of the technical environment, data sources, extraction-transformation-loading (ETL) process, and the robustness of data.
Evaluation of the Physical Environment	PC Terminals	Reviewed the five client computers used to access the HIV disease registry.
	Server	Reviewed the server configuration that is used for the HIV disease registry and contained in the healthcare system's informatics enterprise.
	Cloud Storage	Reviewed the cloud storage environment for the transferring and sharing of patient health data.
Evaluation of the Database	Data Flow	Assessed where and how data are stored and transferred.
	Data Access	Evaluated how data are accessed and the controls involved.
	Relational Tables	Evaluated the relational data tables contained in the Oracle-based database.
	Data Model	Created a logical data model of the HIV disease registry to view how data are organized within relational tables of the registry and describing the relationships between elements.
Database Meetings	Attended Meetings	Attended regularly scheduled quality assurance meetings and management meetings with the HIV disease registry team.
	Meeting Minutes	Reviewed minutes from meetings.

5.4 Limitations and Delimitations

The scope of the case study has several inherent limitations. This study relied on responses from interviewed subjects. Information collected in interviews are subject to recall and inherent biases.

There were also limitations on what data sources could be accessed, which introduces some gaps in knowledge necessary to compare how the data exist in data sources and deriving how data are structured in the HIV disease registry. These restricted data sources included all the data sources owned by the healthcare organization. The case study could only observe data within the HIV disease registry.

The electronic data were HIPAA-protected; therefore, permission from Emory and the healthcare system was required to gain access. Patient data and the HIV disease registry can only be accessed and used within a HIPAA-compliant environment created by the healthcare organization. Therefore, anytime the case study reviewed data it had to be accessed through a client computer located in a health care clinic and organizational policies did not allow data to be saved to electronic devices. There were also organizational policies that constrained the access to the HIV disease registry and patient health data. The process for data access permission is long, complicated, and required approval from multiple departments of both organizations. Additionally, the processes for this were not clearly defined. Access to the healthcare system's data sources was restricted and not permissible. All patient data were accessed through the HIV disease registry or using view-only access of the front-end application of the EHR. This case study had no access to the data or technology infrastructure of the healthcare system to affirm the data contained within the HIV disease registry. The study relied completely on data that were already transferred to the HIV disease registry.

This case study is not a universal model. Data schemas and data relationships are intricate and highly customized to a healthcare system's business and functional requirements so what works for one system is not always applicable to all healthcare systems. The HIV disease registry at the center of this case study may not be the most appropriate solution for other healthcare systems. This case is unique in that it involves a cross-organizational relationship with an academic research institution and a public safety net healthcare system utilizing an Epic EHR accompanied by customized support systems. The case study centers on how data are shared between the two organizations, which is more complex than the technical basis of a disease registry.

In addition to Collaborative Institutional Training Initiative (CITI) certification and Institutional Review Board (IRB) approval, Epic EHR training was necessary to gain access to patient health data. Authentication by the principal investigator was required by the healthcare system to demonstrate the case study was conducted by Emory research personnel. In addition to Emory training also required were signed confidentiality non-disclosure, customer service, and standards of conduct agreements with the healthcare system. The HIPAA security and safety rule policies were examined for both Emory and the healthcare system.

To ensure confidentiality of patient health information all data are stored on HIPAA-compliant Emory Box cloud storage. The cloud storage requires user authorization as well as two-factor authentication. Patient data were only accessible through approved client computers at the healthcare system's facilities. The client computer also requires user authorization and authentication.

6 Results

6.1 Introduction

The HIV disease registry is the result of a multi-disciplinary, cross-organizational collaboration to create an informatics solution that would link an Emory-sponsored disease registry with data owned by a large Southeastern U.S. healthcare system. The registry and research data are managed by the Clinical Research Core of Emory CFAR and Emory LITS. For over 30 years the healthcare system has maintained a Ryan White-funded clinic for specialized HIV ambulatory care that serves the sickest populations of people living with HIV (PLWH). This population is of high interest in public health because it is representative of the U.S. HIV epidemic impacting poor, minority, transgender, and men-that-have-sex-with-men populations. In 2010, the healthcare system launched an electronic health record system (EHR) throughout its main hospital and affiliated community clinics. Since the EHR has matured, there has been increased interest from the major stakeholders in leveraging electronic patient health data for secondary research purposes that contribute to institutional objectives for public health.

CFAR maintains and operates the HIV disease registry that contains data from patient encounters throughout the healthcare system, but primarily includes ambulatory patient data from the specialized HIV clinic and inpatient encounters with PLWH. In addition to having its own informatics team and leadership, CFAR contracts with Emory LITS to manage the data and database. As the information technology hub for Emory University, LITS has the resources to provide the necessary technical knowledge and experience to maintain a large disease registry containing HIPAA-protected health information. The total cost for the development and implementation of the HIV disease registry was $198,581 with an annual operating budget of $152,000.

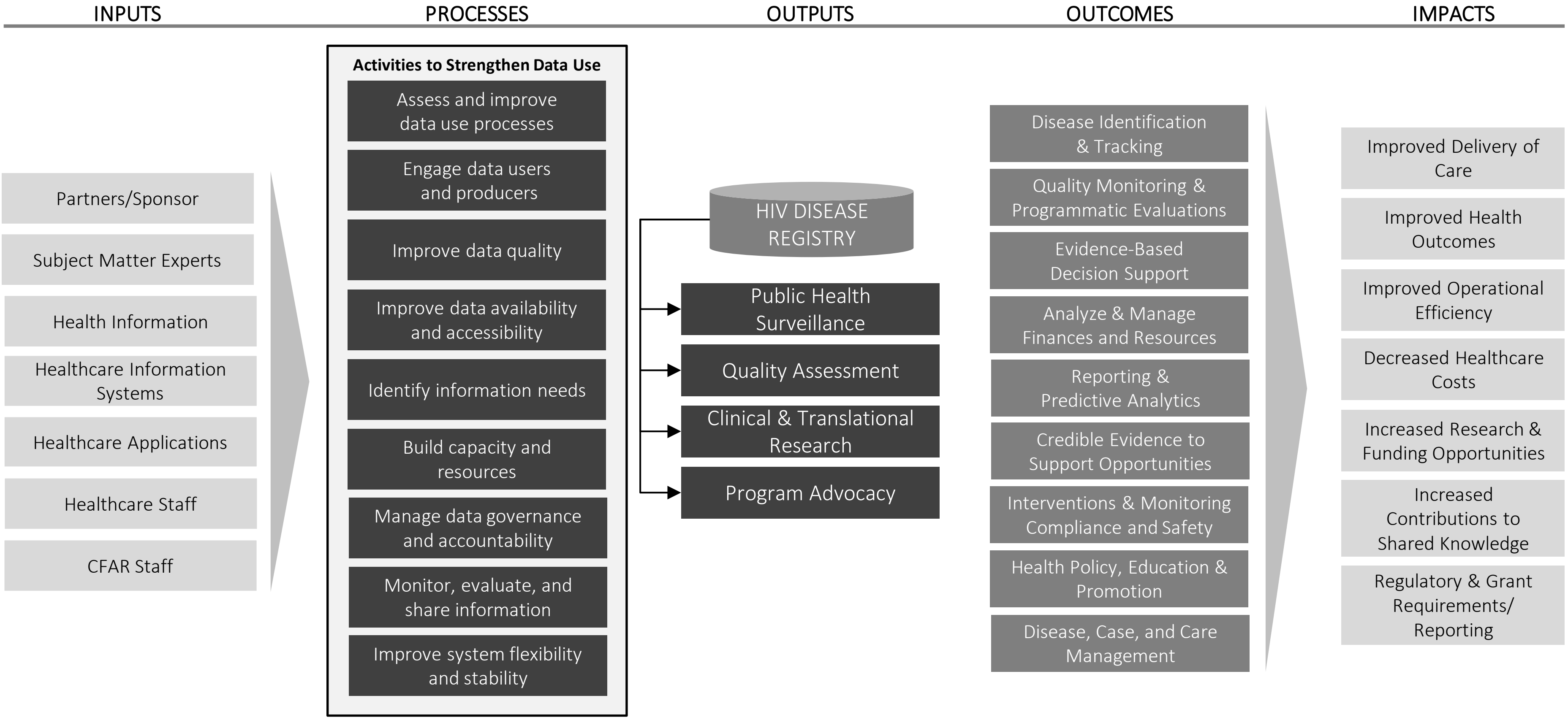
The registry receives annual approval by the Emory University Institutional Review Board (IRB) and the healthcare system's Research Oversight Committee. The protocol used to manage the HIV disease registry has met the requirements and has been issued a full waiver of HIPAA authorization allowing patient health data to be used for research without having to receive consent from each patient. This research protocol was authored by the CFAR informatics leadership to guide the conduct of human-subjects research using protected patient health information collected about individuals receiving care at this healthcare system.

6.2 Program Evaluation Logic Model

The program evaluation logic model of the HIV disease registry illustrates the practical applications for secondary uses of patient health data in a cause and effect linear formation (*Figure 1*). This evaluation model was developed by the case study to diagram how the program operates in applying inputs and processes to achieve outputs that deliver specific outcomes that are necessary to accomplish organizational goals. The processes are positioned as activities that will strengthen how data are used in a disease registry to successfully achieve impacts for improved public health. These processes constitute the foundation for key ouputs of surveillance and research using a disease registry and include primary functions of data and system evaluation that support the outputs. The evaluation model indicates key outputs of a disease registry include public health surveillance, quality assessment, research, and program advocacy that would be functional requirements for any disease or chronic condition. The logic model suggests using patient health data for secondary purposes will achieve outcomes that will produce favorable impacts on healthcare and public health.

In Figure 1 the evaluation model indicates data stewards, staff, systems, applications, and patient data are necessary inputs for disease registries. Processes for data use are developed to enable data availability and accessibility for research. A data use review committee and data use agreements were created to serve the governance requirements for data use. Accountability processes were also implemented to track all uses of patient health data. Iterative collection of business and functional requirements and data quality evaluations are processes that ensure the disease registry meets the data needs of end users. The objectives for the disease registry include public health surveillance, assessment for the quality of care, clinical and translational research, and program advocacy. Program evaluations, reporting, care management, and evidence gathering are outcomes of data uses. Overall, the impacts of the disease registry are improved care and health outcomes, decreased healthcare costs, as well as resulting in increased research and funding opportunities for investigators. These impacts are the results of secondary analyses of patient health data and facilitated by the use of a disease registry.

HIV DISEASE REGISTRY LOGIC MODEL



INPUTS

- Partners/Sponsor
- Subject Matter Experts
- Health Information
- Healthcare Information Systems
- Healthcare Applications
- Healthcare Staff
- CFAR Staff

PROCESSES

Activities to Strengthen Data Use

- Assess and improve data use processes
- Engage data users and producers
- Improve data quality
- Improve data availability and accessibility
- Identify information needs
- Build capacity and resources
- Manage data governance and accountability
- Monitor, evaluate, and share information
- Improve system flexibility and stability

OUTPUTS

HIV DISEASE REGISTRY

- Public Health Surveillance
- Quality Assessment
- Clinical & Translational Research
- Program Advocacy

OUTCOMES

- Disease Identification & Tracking
- Quality Monitoring & Programmatic Evaluations
- Evidence-Based Decision Support
- Analyze & Manage Finances and Resources
- Reporting & Predictive Analytics
- Credible Evidence to Support Opportunities
- Interventions & Monitoring Compliance and Safety
- Health Policy, Education & Promotion
- Disease, Case, and Care Management

IMPACTS

- Improved Delivery of Care
- Improved Health Outcomes
- Improved Operational Efficiency
- Decreased Healthcare Costs
- Increased Research & Funding Opportunities
- Increased Contributions to Shared Knowledge
- Regulatory & Grant Requirements/ Reporting

Figure 1: Program evaluation logic model for the HIV disease registry describing its role for the stakeholders.

6.3 Project Organization

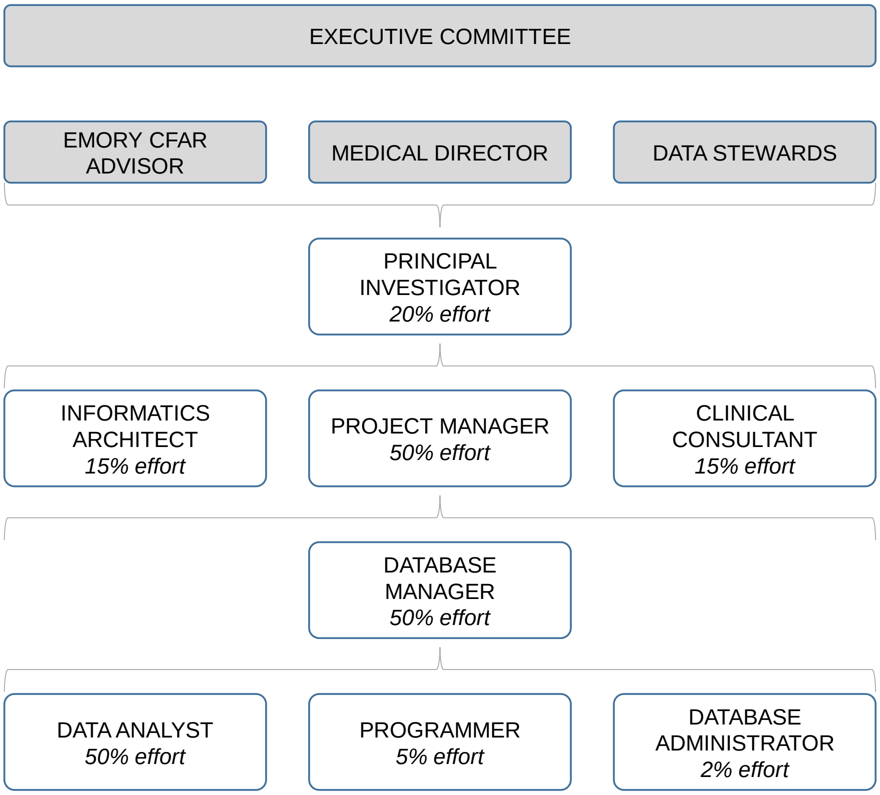


Figure 2: Organizational chart for the HIV disease registry that consists of business and technical units led by a principal investigator with oversight from two committees, an Emory CFAR advisor, the HIV clinic medical director, as well as the data owners.

The HIV disease registry project team consists of a business unit and a technical unit that are both led by the principal investigator (Figure 2). The principal investigator has established an Emory IRB-approved protocol for the management of patient health data for secondary uses in improving quality of care by enhancing clinical processes for people living with HIV who receive care at the HIV clinic. The business unit is comprised with of the principal investigator, Emory CFAR advisor, project manager, and clinical consultant, while the technical unit is led by an informatics architect with a team including a data manager, data analyst, data administrator, and programmer. The roles and responsibilities of all stakeholders including the business and technical units are described in Table 2.

Stakeholder profiles are included in Appendix III: Stakeholder Profiles to help understand requirements to align organizational and business unit strategies.

Table 2: Description of stakeholders' responsibilities in the oversight and management of the HIV disease registry.

Stakeholder Role	Responsibilities
Funding Sponsor, Emory CFAR Directors	Approves project for funding.
Emory CFAR Advisor	Represents the interests of Emory CFAR to monitor progress, management, and procurements.
Managing Sponsor, Executive Committee	High-level oversight of the project as well as leads the strategic direction of the HIV disease registry.
Managing Sponsor, Research Advisory Council	High level oversight of data uses as well as providing subject matter expertise on research applications for the use of patient data.
Technical Advisors	Provides subject matter expertise in research informatics, laboratory informatics, as well as informatics project planning.
Data Steward, EHR	Pass through point for access to healthcare organization's data sources.
Medical Director, HIV Clinic	Pass through point for use of clinic population data; facilitate access to paper records. Acts as the patient advocate ensuring propriety and judicious uses of patient health information.
Data Steward. LIMS	Pass through point for reference laboratory data and subject matter expert for laboratory informatics.
EPIC Clarity Reporting Mgr and Business Intelligence Technical Mgr	Facilitates the transfer of information from healthcare organization's EHR to the HIV disease registry.
Technical Sponsors, Healthcare Organization Information Systems	Provides server space and server support; facilitates the access to data stored in EHR; facilitates access to clarity reports; facilitates access to database templates.
Regulatory Compliance, Emory IRB and Healthcare Organization's Research Oversight Committee	Ensures regulatory compliance to HIPAA as well as standards for respective enterprises; approves research processes.
Principal Investigator	Acts as the leader of the HIV disease registry and its management. Also responsible for creating research protocols and negotiating data uses.
Informatics Architect	Oversees development and implementation of the HIV disease registry; provides subject matter expertise in the technical design and strategy; drives database and software development.
Clinical Consultant (SME)	Physician at the clinic acting as subject matter expert provides knowledge about data used in the care of HIV patients and user interface for the healthcare systems; developed and manages ongoing data strategy.
Database Administrator	Performs backup/restore and database health checks.

Stakeholder Role	Responsibilities
Project Manager	Developed the case study and project management plan for the development of the HIV disease registry; acts as business manager maintaining the schedule, monitoring project progress, designing business workflows and processes, and establishing business reporting mechanisms; acts as liaison between technical team, data stewards, executive committee, research advisory council, and the end users of patient data.
Data Manager	Monitors data performance and quality; involved in data extractions for secondary uses of patient health information.
Data Analyst	Ensures data quality.
Database Administrator	Performs database backup and restoration; monitors database health.
Data Modeler	Designed the database conceptual, logical, and physical data models.
Programmer	Developed data ETL processes and graphic user interface; manages software and ETL processes; manages and designs routine data transfers to migrate data from sources to the HIV disease registry.

Governance Committees: In addition to a project team, there are two committees to oversee the HIV disease registry, the Executive Committee (EC) and the Research Advisory Council (RAC). Establishing these two committees was a key strategy for gaining buy-in from the data owners. The EC includes high-level officials from Emory University and the healthcare organization that included the Deans of the School of Public Health, Global Health, and Clinical Research as well as Emory School of Medicine's Chief Information Officer and the healthcare organization's Chief Compliance Officer. The EC provides high-level oversight and ensures stakeholders are involved in forming the business, technical and research strategies for the HIV disease registry. The RAC plays a different role in governing uses of patient health data extracted from the HIV disease registry. This committee is comprised of patient advocates that include the HIV clinic director, medical director, and physicians that provide care to this patient population. The role of the RAC is to approve any use of patient data for research to ensure judicious uses. This also involves subject matter expertise to help guide the research to ensure that data generated are scientifically robust and publishable.

6.4 Strategic Planning

The strategy developed for the HIV disease registry can serve as a guide for the development of new disease registries. This case study explored the process undertaken by the project team to outline the strategic plan so that it can be shared as a model for the development of a disease registry. The strategies are organized as goals for: determining project feasibility, generating project documents, assessing stakeholder engagement and process alignment, forming an informatics team and advisory groups, and then outlining a project management plan for planning, production and implementation phases. The database itself is a modified instance of a diabetes disease registry, which demonstrates this framework could be translatable to the development of similar databases for other diseases.

Goal 1A – Determine project feasibility

For project feasibility, the key stakeholders will require assurances of access to data sources, informatics tools, and expertise. The project planners will collect both data and technical requirements by examining healthcare data sources and exploring options for a disease registry framework.

Strategy 1A.1: Assess informatics landscape at Emory to determine what tools and expertise are available.

Strategy 1A.2: Assess informatics landscape at the healthcare organization to determine the data sources.

Strategy 1A.3: Determine additional informatics tools and expertise needed for the project.

Strategy 1A.4: Evaluate the physical infrastructure and equipment needs.

Strategy 1A.5: Conduct stakeholder analysis to identify all individuals and groups related to this project; and target their key concerns.

Strategy 1A.6: Generate data use protocols that will require approvals by Emory IRB and the healthcare organization's Research Oversight Committee.

Success Metrics: Identification and successful acquisition of the necessary data sources, infrastructure, and equipment; completion of stakeholder analysis; Emory IRB and healthcare organization approvals for data use protocols.

Goal 1B – Create business documents

Business documents will be tools to gain stakeholder buy-in and will inform project leaders how the disease registry will be managed, secured and governed.

Strategy 1B.1: Gather requirements – business, functional, technical.

Strategy 1B.2: Create business case and project budget

Strategy 1B.3: Create project management plan.

Strategy 1B.4: Create a sustainability plan.

Success Metrics: Completed business case, project management plan, and sustainability plan that are approved by the project leaders.

Goal 1C – Achieve Emory stakeholder buy-in

The project will need sponsorship by Emory School of Medicine (SOM) and Emory CFAR.

Strategy 1C.1: Assess organizational landscape to identify key individuals.

Strategy 1C.2: Present business documents and negotiate partnerships.

Success Metrics: Identification and formation of alliances with the key Emory stakeholders.

Goal 1D – Achieve Healthcare Organization stakeholder buy-in

The project will need sponsorship by the healthcare organization's data stewards.

Strategy 1D.1: Assess organizational landscape to identify processes, necessary documents, and key individuals.

Strategy 1D.2: Include plans for research, governance, security, and propriety to present with business documents and negotiate partnerships.

Success Metrics: Identification and formation of alliances with the key healthcare organization stakeholders.

Goal 1E – Process alignment

The project business and technical functions need to be integrated with existing processes. These functions will also need to meet the requirements of organizational policies that will ensure compliance with federal and state regulations.

Strategy 1E.1: Gather Emory and healthcare organization policies for data, systems, and security. Include policies for HIPAA Security Rule and HIPAA Privacy Rule.

Strategy 1E.2: Gather policies for human subjects research and other regulatory compliance.

Strategy 1E.3: Assess the technosocial environment with Emory and healthcare organization stakeholders.

Success Metrics: Successful formulation of plans for governance, security, and propriety that are integrated with existing processes at Emory and the healthcare organization.

Goal 2A – Access to data sources and resources

Project will need to secure infrastructure and equipment on the healthcare organization's enterprise architecture; access to data elements; and Emory resources.

Strategy 2A.1: Define requirements for data and the system (administrative, technical and physical) as well as needed expertise and regulatory requirements.

Strategy 2A.2: Procure physical infrastructure, equipment, and expertise needs; as well as maintenance, costs, and other requirements.

Strategy 2A.3: Gain access to healthcare organization's data sources.

Strategy 2A.4: Acquire project funding for development and implementation.

Success Metrics: Successful acquisition of data sources and necessary resources.

Goal 2B – Formation of project teams and advisory groups

The HIV disease registry needs individuals to manage the project, govern data uses, and develop/implement the informatics tool.

Strategy 2B.1: Assemble a project management team.

Strategy 2B.2: Assemble an informatics team to develop and implement the registry.

Strategy 2B.3: Assemble an executive advisory council to govern and guide the project.

Strategy 2B.4: Assemble a research advisory committee to govern data uses for research.

Success Metrics: Completed recruitment of all teams/committees.

Goal 2C – Registry Planning Phase

The HIV disease registry will need a plan to identify and acquire specific data elements from the data sources; and will need a plan for a system architecture that is suited for the technosocial environment.

Strategy 2C.1: Define the data and system structures.

Strategy 2C.2: Define the data elements needed from the data sources.

Strategy 2C.3: Develop a strategy to acquire and manage data (ETL).

Strategy 2C.4: Define project costs.

Strategy 2C.5: Define project timeline.

Success Metrics: Successfully created plans and data strategy; as well as cost estimate and timeline.

Goal 2D – Registry Production and Implementation Phase

These are provided in further detail in the Project Management Plan, Section 3.4. Work Breakdown Structure (see Appendix IV: Project Plan).

Strategy 2D.1: Registry design and front-end interface.

Strategy 2D.2: Registry implementation.

Strategy 2D.3: ETL processes design and implementation.

Strategy 2D.4: Post-production.

Strategy 2D.5: Registry maintenance plan.

Strategy 2D.6: Registry sustainability and cost recovery plan.

Success Metrics: The number of projects and users that have accessed data from the HIV disease registry.

6.5 Identification of Patients with HIV

The registry extracts new electronic health data from the EHR each month to identify new HIV cases for inclusion in the HIV disease registry. Once a new case is identified, the entire patient history is exported and appended to the disease registry. The following criteria are used to identify HIV cases from the healthcare system's patient population to be included in the HIV disease registry.

Inclusion Criteria. Patients of any age seen within the healthcare system who have ever had an HIV diagnosis. This was the most direct criteria that provided optimal sensitivity and specificity to include all possible cases. Additionally, any patients having had at least one visit at the healthcare system's specialty HIV clinic were also included even though this clinic provides care to patients that are considered HIV-affected or -indeterminant. These are mostly children of HIV patients seen at this HIV clinic who comprise 2% of the clinic's patient population. It was important to include these non-HIV cases to provide comprehensive evaluations of clinic usage and care delivery.

Exclusion Criteria. Patients with a negative HIV-confirmatory test seen at all clinics within the healthcare system outside of its specialty HIV clinic were excluded as long as there were no subsequent positive HIV-confirmatory tests. Also excluded are patients that have not had a confirmatory HIV test.

6.6 Requirements for the Data

High-level business requirements collected from stakeholder input are included in Table 3 describing the business rules for secondary uses of patient health data. A significant requirement of the healthcare system is that data must be maintained within its informatics enterprise to ensure the data stewards retain optimum security control for where data exists in transmission and storage. These high-level business requirements focus on what Emory CFAR, Emory LITS and the healthcare system needs in order to support their respective missions and goals.

Table 3: High-level business requirements for the HIV disease registry.

1	System will be maintained within the healthcare system enterprise on their servers behind their firewall.
2	System will facilitate access to aggregated data from disparate data sources that will include EPIC, LIMS, and pharmacy systems, as well as clarity and workbench reports, in addition to other medical records associated with healthcare system patients including paper medical charts.
3	System will support clinical & translational research in public health. This includes access to data and the ability to export data sets to perform analysis.
4	System will support quality of care and performance metrics. This includes access to data and the ability to export data sets to perform analysis.
5	System will facilitate program evaluations.
6	System will include a user interface that allows for manual entry of data.
7	System will include a front end user interface that will support database management, querying, and extraction of sample data sets.
8	System and personnel will follow applicable requirements of the HIPAA Security Rule, HIPAA Safety Rule, and any appropriate Emory or healthcare system SOPs that define physical, technical, administrative requirements, accountability, and data use governance.
9	System will include a sustainability plan for ongoing maintenance, support, and funding.

6.7 Description of the Data

The majority of the electronic data collected in the HIV disease registry are extracted from the Epic EHR system (Madison, Wisconsin) utilized by the healthcare system. The case study found that as of July 2018 the data include 13,033 unduplicated outpatients since 2010 and 13,538 unduplicated inpatients representing 80,775 human years. There were additional legacy data manually loaded into the EHR going back to 2000 that were captured by the disease registry. The HIV disease registry also includes 348,731 CD4 and viral load lab results that are key markers for HIV disease progression and 540,143 patient encounters. Data are extracted and loaded into the HIV disease registry from data sources that include the EHR (Epic), the currently active laboratory information management system (LIMS), a pharmacy database that includes dates and times of medication orders and pickups (RX30), as well as laboratory results data prior to EHR implementation from the legacy LIMS (Ultra C) (*Figure 3*). At this time, only structured data are available through the HIV disease registry. In addition, some data have recently been extracted from the Epic workbench reporting tool to be included in the HIV disease registry. All data are stored within the healthcare system's informatics enterprise to allow the data stewards to maintain control of the electronic patient health data contained in the HIV disease registry.

There are 176 covariates extracted from the data sources for the HIV disease registry. These variables were strategically selected based on published guidelines for HIV healthcare performance metrics. In addition, some variables were selected for current and potential trends in research. This provides scalability for future studies in which to better understand the potential explanatory factors that impact HIV health outcomes. The variables were selected by physicians, informaticians, and biostatisticians that were consulted during the development of the registry.

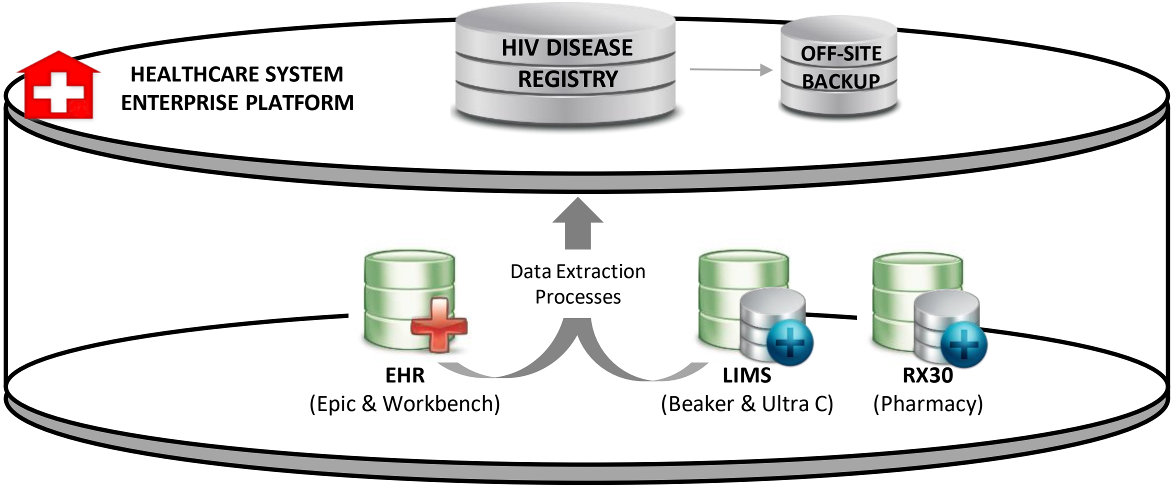


Figure 3: Sources of data include healthcare applications used by the healthcare system.

The routine data extractions are automated to extract, transform and load (ETL) electronic patient health data to a file transfer protocol (FTP) server provided by the healthcare system serving as a staging area. These automated data transfers are facilitated by standardized SQL scripts written by a

LITS ETL-programmer. Nine SQL scripts are involved in the migration of data from the healthcare system's data sources to the HIV disease registry:

1. Identify patients having an HIV diagnosis
2. Identify patients having a visit at the HIV clinic
3. Extract outpatient visits, diagnoses, problems, and vitals data
4. Extract inpatient admissions, diagnoses, problems, vitals, and mortality data
5. Extract prescribed medication data
6. Extract lab results data
7. Extract procedure data
8. Package data into relational data tables

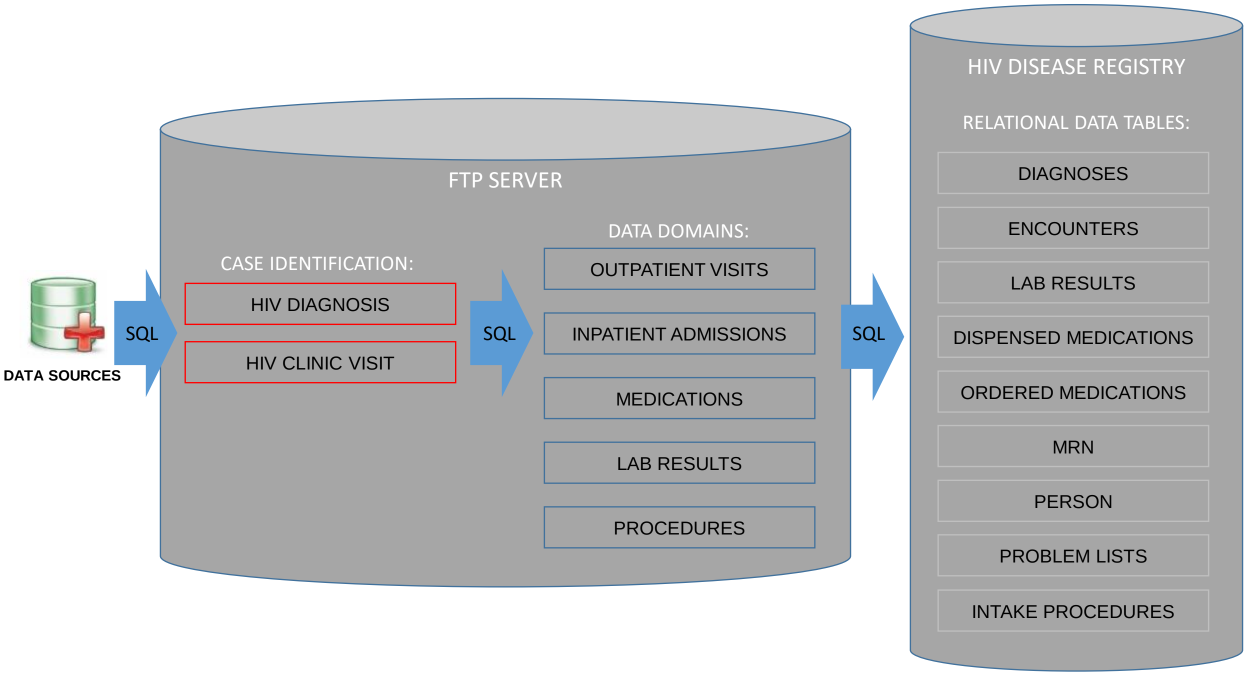
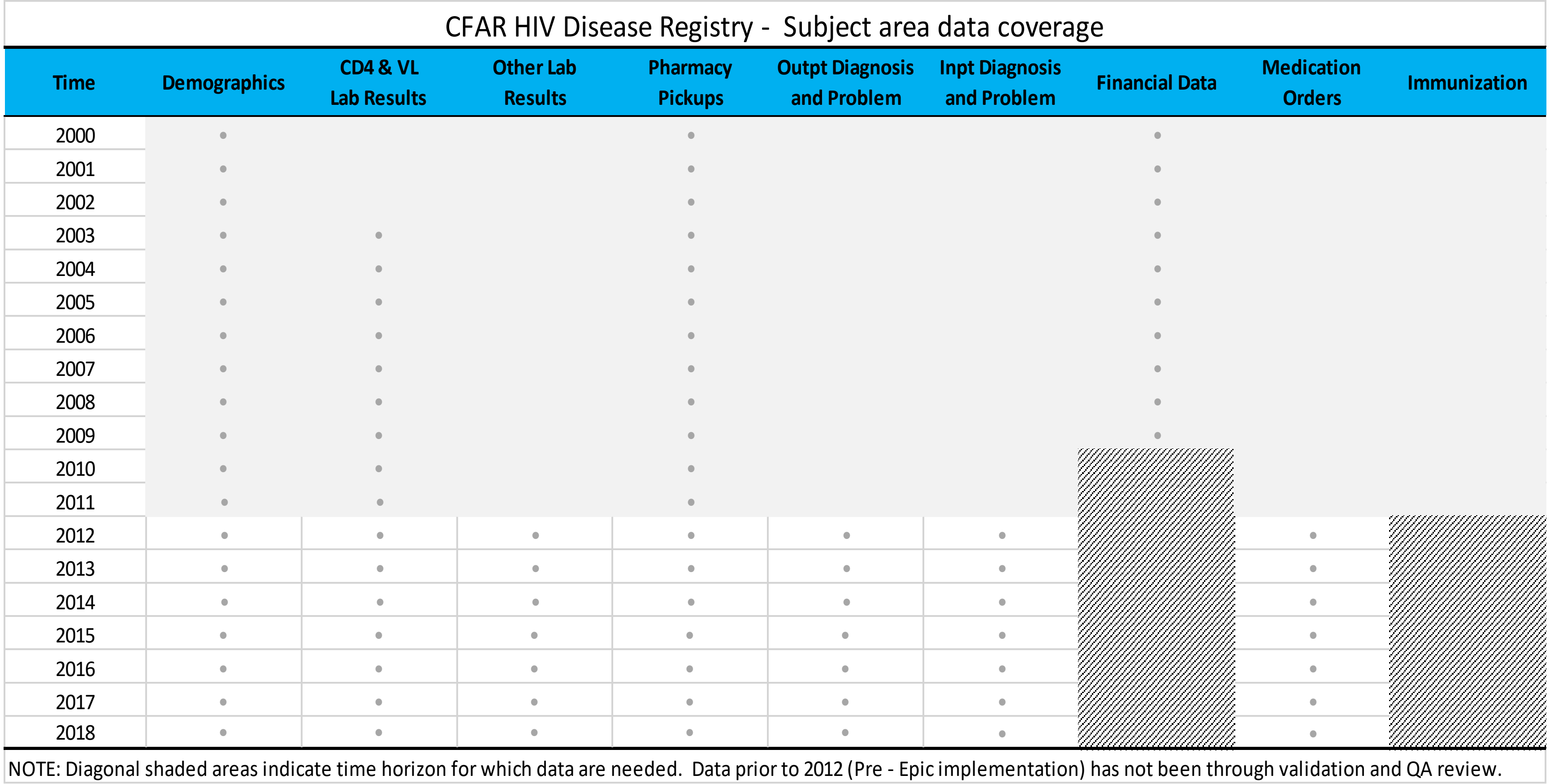


Figure 4: Extraction, transfer and loading of data from the healthcare system data sources to the HIV disease registry are facilitated by using customized SQL programs.

The first two SQL scripts identify patients having an HIV diagnosis and patients having a visit at the HIV specialty clinic. Once the cases are identified, the associated clinical data are extracted into distinct domains. Each domain has a unique SQL script to automate the data extraction and transfer: outpatient visits, inpatient admissions, prescribed medications, lab results, and procedures. These domains of data are maintained within the FTP workspace where the data can further be organized into the final relational tables that will exist in the HIV disease registry (*Figure 3*). Demographics and pharmacy pickups are the most comprehensive data type that includes information from 2000 to 2018. CD4 and viral load lab results go back as far as 2003. Financial data are only available for 2000 to 2010. All other data, including other lab results, outpatient diagnosis and problem, inpatient diagnosis and problem, and medication orders are available from 2012 to 2018 (Table 4).

Table 4: Periodicity of electronic patient health data included in the HIV disease registry range from 2000 to present. Source: Emory LITS, author Jeselyn Rhodes.



CFAR HIV Disease Registry - Subject area data coverage

Time	Demographics	CD4 & VL Lab Results	Other Lab Results	Pharmacy Pickups	Outpt Diagnosis and Problem	Inpt Diagnosis and Problem	Financial Data	Medication Orders	Immunization
2000	●			●			●		
2001	●			●			●		
2002	●			●			●		
2003	●	●		●			●		
2004	●	●		●			●		
2005	●	●		●			●		
2006	●	●		●			●		
2007	●	●		●			●		
2008	●	●		●			●		
2009	●	●		●			●		
2010	●	●		●			▨		
2011	●	●		●			▨		
2012	●	●	●	●	●	●	▨	●	▨
2013	●	●	●	●	●	●	▨	●	▨
2014	●	●	●	●	●	●	▨	●	▨
2015	●	●	●	●	●	●	▨	●	▨
2016	●	●	●	●	●	●	▨	●	▨
2017	●	●	●	●	●	●	▨	●	▨
2018	●	●	●	●	●	●	▨	●	▨

NOTE: Diagonal shaded areas indicate time horizon for which data are needed. Data prior to 2012 (Pre - Epic implementation) has not been through validation and QA review.

New data are integrated by appending the newly identified data to the existing relational data tables organized within the HIV disease registry. The nine relational tables include data about diagnoses, clinical encounters, lab test results, medication dispensations, medication orders, medical record information, demographics and contact information, patients' problem lists, and procedures ordered at clinical encounters (*Table 5*).

The EHR validation process is performed by an Emory LITS data analyst, Emory CFAR analyst and an infectious disease physician that serves as a clinical subject matter expert. Once data are extracted from the EHR, the three reviewers conduct a data quality evaluation to determine data completeness, consistency and accuracy. The evaluation consists of individual data and aggregate data analyses for qualitative assessment of these data characteristics (Appendix V: Data Quality Plan). Completeness is the presence of required data, accuracy is the closeness of agreement between a data value and the true value, and consistency is related to precision in which the relevant uniformity in data are measured [40]. This evaluation involved selecting a random group of cases (n=20 to 200 cases) to validate the data using various aspects of data collected within the EHR that included laboratory results, progress notes, problem lists, diagnoses, and respective clinical dates. With a normal distribution, the central limit theory holds that a sample population of $n \geq 30$ is enough to be representative of the study population when variances are controlled. In addition to comparing EHR data with extracted data, the reviewers also use Epic workbench data sets for comparison. Workbench is a data reporting tool included in the EHR that captures real-time, cross-sectional data reports that can be readily accessed by health care providers but is limited in that it cannot perform longitudinal analyses. Limiting workbench from doing longitudinal analyses ensures there is no strain on applications used during the delivery of care. Workbench functions as an operational tool to assess patient care and quality. The Epic clarity reporting tool, which serves more of an analytical function, has also been used for data validation. These tools use the same data source as the HIV disease registry and are useful for ensuring data is properly translated during migration from the data source to the disease registry. Clarity does not provide data that is real-time, typically running 1 to 2 weeks later than workbench.

Table 5: Domains of electronic patient health information extracted from healthcare system data sources and migrated to the HIV disease registry.

Domain Name	Description
DIAGNOSIS	Records of diagnosis as recorded in relation to outpatient and inpatient encounters within the healthcare system
ENCOUNTER	Records of outpatient and inpatient encounters within the healthcare system
LAB	Records of laboratory results; only specific labs are captured
MED_DISPENSED	Dispensed medications from the specialized HIV clinic's pharmacy pulled from the pharmacy database
MED_ORDER	Records of medications ordered (prescribed) including provider identification, dosage, and quantity prescribed; includes medications ordered in outpatient and inpatient settings
MRN_AKA	Alternate medical record numbers reconciled with the master medication record number within the healthcare system's Epic instance
PERSON	Record of patients in the registry; data contains demographics, address, and insurance coverage
PROBLEM_LIST	Records of patients' problem list in Epic with status indicator ('Active'-'Resolved'-'Deleted')
PROCEDURE_INTAKE	Records of procedures ordered at clinical encounters

6.8 Description of the Database & Data Storage

The HIV disease registry exists in a complex technical landscape because it contains protected patient health information that is not owned by Emory CFAR. The registry is in a shared cross-organizational environment in which it resides within the healthcare system's informatics enterprise where the healthcare system administrators can maintain administrative control while the relational database is managed by Emory entities. The shared framework allows the registry to employ its own hardware, a blade server (Dell PowerEdge with 2.8 ghz Intel Xeon processor E7), directly onto the healthcare system's existing enterprise architecture. This facilitates the healthcare system's oversight for how data are shared. It also allows the healthcare system to maintain data backups on its own servers so that protected health information remains contained within the healthcare system's enterprise and meets the organization's security protocol requirements.

The exception to the HIV disease registry's data existing on the healthcare system's enterprise is when datasets are extracted for research, in which case it is migrated from the healthcare system's enterprise to an Emory-sponsored cloud-based storage. The healthcare system lacks its own secure platform to facilitate HIPAA-compliant sharing of protected patient health information. Emory maintains licenses with Box, a HIPAA-compliant cloud storage provider, for which the healthcare system has agreed to use as a tool for sharing data externally with investigators for secondary research. Emory CFAR also purchased off-the-shelf applications that are used by Emory CFAR and Emory LITS client PCs to interface with the data contained within the registry. These applications include Dell Systems Toad for Oracle Base Edition v12. 6, a relational database management tool, and Oracle Instant Client v11. 2.0.4 that enables remote connection and communication with the Oracle database.

Software for the HIV Disease Registry

- Oracle SQL Database 11g (available at no cost through Oracle)
- Oracle SQL Developer Data Modeler v3.1.1.703 (available at no cost through Oracle)
- Oracle SQL Developer v3.1.07 (available at no cost through Oracle)
- Oracle SQL Instant Client v11.2.0.4 (available at no cost through Oracle)
- SAS statistical analysis software ($200/year for group license through Emory University)
- Dell Systems Toad for Oracle Base Edition v12.6 (one time license $1,200 per seat)

The data for the HIV disease registry are extracted from its sources and then aggregated in several places: the staging area where ETL processes occur, within the HIV disease registry itself, or in data sets that are extracted for approved data uses. Staged data are housed on mandatory HIPAA-compliant FTP storage solutions provided by the healthcare system. Queries are run on the database by an authorized data analyst to extract only the minimum information necessary. To ensure patient confidentiality the data extractions are managed by an Emory honest broker system that de-identifies and anonymizes patient health data using a coding system for which a key to link the data is produced and managed by the Emory honest broker. The key is not shared with data users. Authorization and authentication protocols are both used for secure access to patient health data. Authorization roles are used to determine the level of data access, which are controlled by Emory LITS, the certified honest broker at Emory. Data users that can demonstrate a need for identifiable patient data are required to have a protocol for human subjects research that is approved by both Emory IRB and the healthcare system's Research Oversight Committee.

Technical Landscape

The systems and their data content are organized as a single, centralized data storage with management configuration to maximize data integrity and minimize data redundancy. Centralizing the data removes it from the healthcare system's workflow so that it does not impede on clinical performance, data quality or accuracy. The assumption was this single source contained all the necessary data to act as a disease registry that contains scientifically robust data that can effectively be used in secondary analysis. Epic is a fixed, proprietary architecture that extracts data from various sources to be consolidated within a single framework that supports patient care and billing using an object-oriented database management system with hierarchically structured data. The design and functionality support transactional data constrained by a single governing schema. The HIV disease registry offers active data processing to better support secondary uses and can be a flexible platform that can consume, aggregate, transform, and enrich data for both business accuracy and scientific research. Its structure is formulated as a client-server model for distributed partition of tasks between client and server hardware.

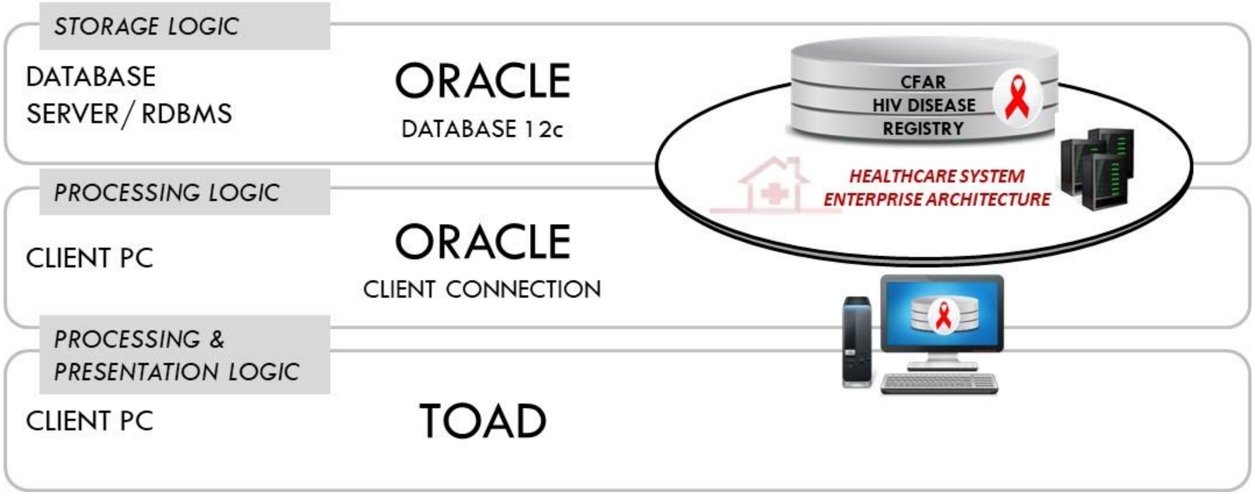


Figure 5: The storage, processing and presentation logic of the HIV disease registry.

The client-server system constitutes three components that include the storage, processing and presentation logics (*Figure 5*). In the HIV disease registry framework these are based on a two-tier database server architecture where the client is responsible for the processing and presentation logics while the storage logic is restricted to the server. The client is a user-accessible workstation PC that is used to request a service from the server. The server is a PC that is designed as a mainframe to provide services from the HIV disease registry database. The server also comprises the physical location for the HIV disease registry's database. The storage logic represents database management system activities such as data storage and retrieval. For the HIV disease registry, this storage logic is the Oracle database. The processing logic includes how data are accessed, the business rules, and data management constructs for procedures and functions. The processing logic is primarily confined to the client PC. The presentation logic represents the graphic user interface accessed by using the client PC via the Toad application. This encompasses the user inputs and outputs. The partitioning of applications across these three logics provides improved performance, improved interoperability, balanced workloads, and enhanced accountability for the client-server architecture.

Security

A detailed description of the data security plan is included in Appendix VI: Data Security Plan. The HIV disease registry informatics team can access this data from a physical location that is connected to the healthcare system's enterprise network server. The team can also remotely access this data through a process managed by the healthcare system that only authorizes select individuals. The Emory security controls have also adopted secure two-factor authentication and role-based credentialing processes with permission controls to access Emory resources and its virtual private network (VPN) connection.

The registry team has access to HIPAA-compliant servers located within the healthcare system and Emory SOM that employ encryption mechanisms to ensure the security of sensitive data at rest and in transit. Messaging and data transfer is modulated within a virtual private network through user authorization and authentication protocols. Utilizing these informatics tools offers enhanced security and accountability with storage and transfer encryption. The HIV disease registry has adopted both the healthcare system and Emory security and HIPAA-compliance standard operating procedures that defines requirements for technical, physical, and administrative controls.

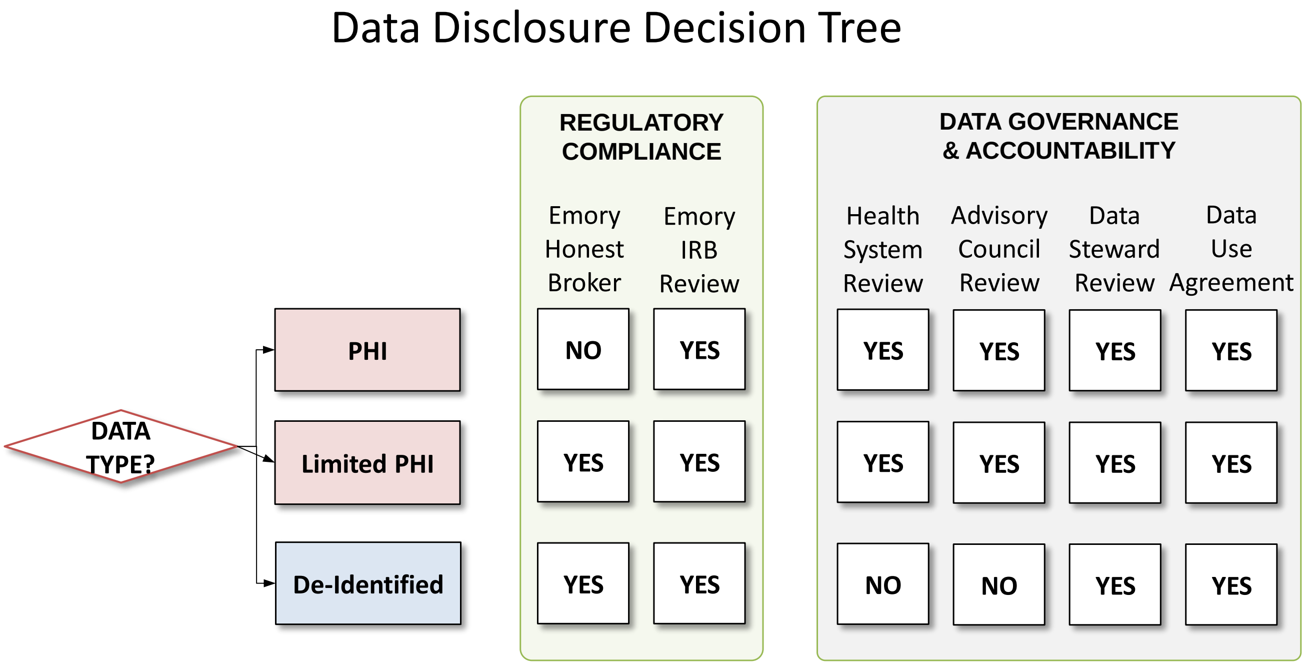
6.9 Governance of Data Sharing

An advisory research council was created to offer oversight for the sharing of data in secondary analyses. The council consists of data stewards from the healthcare system, medical directors and physicians that are involved in the care and treatment of this patient population. This was an important step in the development of the disease registry to ensure stakeholders remain involved in the decisions to share data and continue to have buy-in. All requests to release data from the HIV disease registry is reviewed and approved by this council for the scientific merit of the intended use of patient health data.

Two different applications of data sharing are encountered when using patient health data from the HIV disease registry for secondary purposes (*Figure 5*). Data are used for research purposes and non-research purposes that are submitted to an advisory research council as part of the accountability process to determine what data can be released, the intended purpose of the data, identities of all the data users, and where data are allowed to be stored. Non-research data uses are not required to be submitted for compliance to the Emory IRB because data are used for operational and performance purposes without intent to publish or share data. These uses can include HIPAA-protected personal health information (PHI), limited PHI or de-identified data, however, regardless of the type of data used for non-research purposes all types of data are reviewed under the same process before data are released. The governance process for non-research uses requires approval by both a CFAR director and the healthcare system's assigned data steward.

The use of patient health data for research purposes is more complex because of HIPAA Security and Privacy Rules' requirements for the protection of human subjects' data involved in research intended to be published or shared. Figure 5 depicts the paths for the sharing of PHI, limited PHI or de-identified patient health data. All types of data are subject to processes for data governance and accountability that requires a formal data request by the user of data describing intended usage, approval from the research council, the data stewards, as well as a signed data use and data destruction contract. Release of PHI and limited PHI requires additional approvals from Emory IRB and the healthcare system's Research Oversight Committee. Any de-identification of data are controlled through the Emory Honest Broker to ensure that de-identification and anonymization of data meets HIPAA compliance in accordance with Safe Harbor rules that provide guidance on how to apply the HIPAA Privacy Rule by removing information that can be used to identify patients.

In addition to council review of data requests, data sharing is governed by several data agreements that include a HIPAA business associate agreement (BAA) between Emory University and the healthcare system, a memorandum of understanding (MOU) between Emory CFAR and the healthcare system, and data use agreements (DUA) between Emory CFAR and users of data. Three DUAs are in place to manage the sharing of PHI, limited PHI, and de-identified data products with end-users with versions for Emory users and non-Emory users. The DUAs and MOU were designed by legal associates from Emory Office of Technology Transfer (OTT) and the Program Manager of Emory CFAR Clinical Research Core.



Data Disclosure Decision Tree

DATA TYPE?	REGULATORY COMPLIANCE		DATA GOVERNANCE & ACCOUNTABILITY			
	Emory Honest Broker	Emory IRB Review	Health System Review	Advisory Council Review	Data Steward Review	Data Use Agreement
PHI	NO	YES	YES	YES	YES	YES
Limited PHI	YES	YES	YES	YES	YES	YES
De-Identified	YES	YES	NO	NO	YES	YES

Figure 6: Decision tree describing how data from the HIV disease registry are shared with end users. This figure is adapted from Emory LITS and modified to conform to the Emory CFAR HIV disease registry.

The BAA is a high-level agreement between the two institutions that is intended to cover all Emory entities including Emory CFAR. The BAA applies to PHI, including limited data sets, and is not intended for governing the sharing of properly de-identified data. The agreement asserts that the healthcare system has to approve any disclosure of PHI and any de-identified data that mistakenly includes PHI would have to be reported as a data breach. Highlights from the BAA include:

6.9.1 BAA Use and Disclosure of PHI

- Data can be shared with third parties if the third party assures data confidentiality and agrees to promptly notify Emory CFAR of breaches.
- The healthcare system has the sole responsibility to approve PHI disclosure/use. It is Emory CFAR policy to require approvals or exemption from the healthcare system research oversight

committee and Emory IRB for any data request that includes PHI. However, in the meeting notes between Emory CFAR and Emory OTT the case study noted there is uncertainty from Emory OTT legal representatives about whether this approval is sufficient for the healthcare system's office of compliance.

6.9.2 BAA Safeguarding and Reporting Misuse

- Emory CFAR will use safeguards to protect data; and will mitigate disclosures. The case study observed Emory CFAR follows the HIPAA Security Rule policies of both Emory University and the healthcare system. As required by Emory University policies, Emory CFAR uses Emory LITS as the honest broker for proper de-identification and anonymization of PHI shared with end-users using HIPAA-required Safe Harbor methodology defined in 45 CFR 164.514b(2) by the U.S. Department of Health and Human Services. Emory CFAR also employs Emory University policies for technical, administrative and physical safeguards for accessing and managing PHI that are more stringent and detailed than what the case study observed in the healthcare system's HIPAA Security and Safety policies.

- Emory CFAR must notify the healthcare system of any breach within five business days of becoming aware. However, the review of this case study finds the agreement does not define who at the healthcare system should be notified. It is also Emory University's policy that Emory CFAR will report data breaches to the Privacy Officer at the Office of Research Compliance as well as the Emory IRB.

6.9.3 BAA Access to PHI

- Emory CFAR agrees to furnish the healthcare system with PHI maintained in the disease registry when requested.

- A request to Emory CFAR that includes PHI in the disease registry must be forwarded to the healthcare system within five business days. The decision to release data for this request is the healthcare system's sole responsibility. The case study determined the intent of this is for an individual requesting their own information, but this agreement does not cover situations like Emory CFAR acting as a data distributor. There is uncertainty from Emory OTT legal representatives about whether Emory CFAR could be defined by HIPAA as a clearinghouse. If so, this situation may require an additional BAA between Emory CFAR and the healthcare system. To resolve this Emory CFAR is working with the healthcare system office of compliance to establish a workflow to approve requests for data from the HIV disease registry.

6.9.4 BAA Accountability of Data Uses

- Emory CFAR will document any PHI disclosure. The case study determined the intent is to require Emory CFAR to maintain accountability for any PHI release.
- The healthcare system must approve any requests for accountability of data usage; Emory CFAR must forward requests for data within 5 days. The case study determined the intent of the language in this section is to inform the healthcare system of any formal audits or evaluations of the HIV disease registry.
- Additionally, the BAA does not appear to have any issues with using PHI for non-research uses. Non-research uses are operational uses intended for quality initiatives or evaluation that inform decision making for the management of the clinic or its patients. In contrast to the BAA, the case study observed that Emory CFAR has a defined process for this. Emory CFAR asserts these types of non-research uses should be endorsed and led by a senior clinic administrator or medical director. Language to define this process was specifically included in the MOU to address how data are shared when not used for research purposes.

The MOU has been in active negotiation between Emory CFAR and the healthcare system as the two organizations work to define how data can be shared. This arrangement allows Emory CFAR to extract data from the healthcare system's medical records for secondary uses that will support public health and clinical research. The healthcare system retains ownership of their medical record data and Emory CFAR/ Emory University do not acquire ownership simply through being contracted to perform a data extraction for research uses. The case study determined that further use of any of this data either needs to be approved by the data owner or the healthcare system through their execution of appropriate data transfer agreements; or they need to transfer authority to Emory CFAR to execute those agreements and act as a broker in sharing data on behalf of the healthcare system.

However, there is uncertainty about whether the BAA would cover the brokerage of data by Emory CFAR. If Emory CFAR acts as a broker of this data, they would likely be a business associate of the healthcare system and would thus need to execute a new BAA between Emory University and the healthcare system. This BAA could also lay out what authority Emory CFAR will have to use and disclose the data on behalf of the healthcare system, and whether Emory CFAR can execute data transfer agreements for the data without involvement of the data owner, which is also covered in the MOU.

For a legally executed data use agreement, Emory CFAR would specifically need an Emory University designated signatory given authority as a representative of Emory University to enter into these types of legal agreements. Emory CFAR does not have such a designated signatory; therefore in addition to establishing the MOU between Emory CFAR and the healthcare system, Emory CFAR will need to negotiate with Emory University to determine if there can be a suitable designated signatory within Emory CFAR that can enter into legal contracts (DUAs).

In addition to an institutional level signature, it would be prudent to require additional business unit signatures from a designated signatory from Emory CFAR, the healthcare system's HIV clinic, and from the medical director responsible for the welfare of these patients. The case study observed that language for this is incorporated in the MOU; however, it does not identify the designated business unit signatories. An Emory institutional level signature will not be necessary if the healthcare system provides its own DUA between the healthcare system and the end-user/ end-user's institution. The healthcare system has not provided these necessary documents and guidelines, but Emory CFAR is still negotiating this process with the data owners. It will be difficult to achieve an uncomplicated, efficient process for this, but should be among the chief priorities in negotiating with the healthcare system to provide a clear path with designated authorities.

7 DISCUSSION

7.1 SUMMARY OF THE CASE STUDY

This descriptive case study uses a multi-modal approach that involves a variety of methods including interviews, field examination, protocol and document analyses, direct observations, archived document review, and exploration of the physical infrastructure. The case describes the successful exchange of a healthcare system's protected electronic patient health information with Emory CFAR in which the data are used to populate an HIV disease registry that can facilitate secondary analyses. Historically, this patient data was difficult or impossible to obtain by Emory CFAR investigators. The HIV disease registry has opened opportunities for Emory CFAR investigators to explore public health about a large population that is representative of the U.S. HIV epidemic.

The purpose of this case study was to explore synergistic uses of patient health data between the two organizations. The case study intends to share this knowledge with investigators developing new registries targeting other diseases. In this case study, the ecosystem of the HIV disease registry was reviewed in detail to gain insight into how data are collected, managed and shared by Emory CFAR. The results reveal the governance and accountability requirements provide the foundation of trust for the use of patient data by an external organization. These findings describe the creation of a relational database based on a model that organizes data from the EHR in a format that is appropriate for research. The registry is among the first collaborative disease registries developed in this cross-organizational environment, demonstrating that patient data could safely and successfully be used for secondary analyses.

While case studies typically lack broad generalizability, the amount of detail with this case study on the Emory CFAR HIV disease registry can still provide insight on a common issue surrounding judicious sharing of patient health data. Investigators of other diseases and chronic conditions also utilize patient health data for secondary analysis and would be interested in learning how others have successfully driven data research by using information derived from patient care. Using this case study, investigators can apply this knowledge to their environments, even if the HIV disease registry is not exactly applicable to their own situation. While the specifics of the HIV informatics solution are not universal, the information can be translated into strategies for replicating the disease registry with any health condition or disease, which would use the same EHR-linked data source. In this case study, the framework consisted of strategies for interoperability, governance, accountability, and security. These are the common themes for any disease registry that is being developed, especially those that are established in an environment in which the investigators do not own the patient health data. These are important strategies for understanding and aligning organizational cultures in order to overcome barriers to data accessibility.

7.2 Implications

The case study contributes information to improve data accessibility for public health research using a disease registry. This study presents a solution for how data can be accessed for research as well as improving the time it takes to access data. The case study itself is a framework to help bridge the communication between research, business and technology. Using this case study prepares the blueprints necessary for developing a disease registry that respects propriety for using protected health information. The HIV disease registry in this case study demonstrates that inter-institutional data sharing can be accomplished by applying thoughtful strategies for governance, accountability and security. The case study establishes that these strategies were instrumental in proving to data stewards that the HIV disease registry's project team considered how it should conduct judicious uses of patient health data.

The healthcare system is one of the largest safety-net hospitals in the Southeast and manages a vast amount of health data; however, access to recent population health data for research is limited and not available in a timely manner. The barriers for accessing patient health data are complicated for academic investigators using data about this patient population owned by the healthcare organization. Without research support or a data dictionary provided by the healthcare organization, investigators will have to understand data structure in the EHR without providing investigators with a data dictionary. Another barrier is that investigators often do not include a data strategy in the early stages of research development. The HIV disease registry consists of a team of clinical and data experts that can bridge the communication between research and technology. These health data are necessary to understand the demographic and clinical characteristics that impact people living with HIV as well as the social and environmental factors that can influence this population's health outcomes. The knowledge gained from research can be used to target and evaluate interventions to improve population health. Lack of access to comprehensive, real-time data can limit research that can advance the evidence-based delivery of care, determine effective allocation of resources, and contribute to the body of scientific knowledge. This case study can provide a template for developing new disease registries in this healthcare system that can target other diseases such as reproductive health, pulmonary disease, cancer, or diabetes. The case study can also be applied to other organizations with a similar environment of inter-institutional data sharing.

A key feature of the HIV disease registry is its relatively low development and implementation costs. The development and implementation cost of $200K and annual operating cost of $152K ordain this disease registry as a cost-effective tool that returns great value to public health. Projects such as this are often scoped at ten times this cost. The project archives revealed that the first proposal for the HIV disease registry was projected at $3M with an annual operating cost of $1M, which was rejected by Emory CFAR directors. The project leadership then turned to Emory LITS and with endorsement from the Emory SOM chief compliance officer was able to use a departmental informatics architect to develop a new plan that leveraged existing resources at Emory SOM, Emory CFAR and the healthcare organization. The architect was able to use an instance of an existing diabetes database and conserve the schema to build the framework for the HIV disease registry. Building the HIV disease registry from the diabetes disease registry is an example of translating an existing disease registry for another disease. The HIV disease registry employs a basic, low-cost framework capable of supporting patient health data for other diseases.

Contributing to the successful implementation of the HIV disease registry was the project leadership achieving stakeholder acceptability through strategically planned engagement that resulted in confidence in the project and buy-in of the healthcare organization's data stewards. In the early planning phase, the project managers had identified key champions at Emory who included Emory CFAR leadership and Emory SOM's chief information officer, as well as key champions at the healthcare organization. Identifying those key champions at the healthcare organization was a particular challenge. The project managers targeted the three gatekeepers to accessing patient health data: the healthcare organization's information technology, compliance and quality departments. The initial attempt at gaining buy-in from the chief information officer did not present an opportunity to engage with them to establish a disease registry. The IT department was devoting resources to a relatively new and maturing EHR system and could not commit the necessary bandwidth to support the development and implementation of the HIV disease registry. The project team encountered its champion during the process of gaining research approval for secondary uses of patient health data. The healthcare organization's chief compliance officer (CCO) ascertained the HIV disease registry as an opportunity to advance the departmental goals for enhancing clinical care through secondary uses of patient health data. The CCO credited the plan for the HIV disease registry that included strategies for governance, accountability, security, and interoperability as an ideal template for judicious sharing of protected health information with its external partners. The CCO provided sponsorship for the HIV disease registry and brought in the chief quality officer to corroborate in using the registry to help shape the healthcare organization's policies and procedures for sharing data with external partners.

The HIV disease registry is able to communicate nearly real-time population health data. Before the introduction of this, registry investigators had to request data directly through the healthcare system that had a complex, long and uncertain process. Some trial and error for an investigator was needed to eventually receive sufficient data necessary to conduct research. So, there may be several data request submissions as their research evolves, adding more time needed to gain access to data. Many times the use of data for public health research was approved, but placed in a long queue for extraction by the health care system. An interview with the healthcare system's chief compliance officer provided insight to how the resources used to support these types of secondary analysis did not exist, so there was competition to access existing data analysts, which levied a strain on the healthcare operations. The existing resources were not intended to support secondary research and were established before the voluminous amount of electronic data; therefore, the path to share data was unchartered and untested. Furthermore, existing resources were prioritized for hospital requirements above the needs of research, so even a promised date of data delivery could be unmet if there were competing needs of the hospital. This uncertainty of gaining access to recent health data is detrimental to the petitioning of grant funding which is typically defined by strict deadlines for submissions to be considered. This case study describes how these issues were resolved and provides insight into how protected data are shared between the two institutions, particularly the roles and responsibilities in data governance, accountability and security.

The case study describes the strategies that were formulated to foster a relationship for inter-institutional data sharing. After five years of planning and negotiating with the healthcare system, Emory CFAR developed strategies for data governance, accountability and security that would ensure the healthcare system's buy-in to the implementation of a disease registry. These three strategies helped to assuage the key concerns a healthcare system would have in sharing protected electronic patient health data with an external entity. To establish a disease registry similar to Emory CFAR, an academic research organization has to overcome the barriers of the healthcare system: (1) having the resources and skills to support secondary research using patient health data, (2) limiting the strain on the existing system so that there is no impact on healthcare business, (3) providing ongoing funding to ensure continuity for a disease registry, (4) creating an infrastructure for overseeing and tracking all uses of data, (5) developing an amenable plan for sharing data that offered propriety and security, and (6) keeping the data secure. These concerns are common and translatable barriers to any healthcare system sharing data externally with an academic institution and provides the framework for a suitable strategy for data sharing between organizations.

Providing Knowledge, Skills, and Experience

Firstly, a barrier for the healthcare system is lack of resources and infrastructure for supporting data research. However, Emory CFAR and Emory LITS brought experience with data research and management. Emory CFAR has been doing research with patient data since 1998. Emory LITS staff are trained and experienced in broad optimizations for secondary uses of patient health data because it has managed these kinds of data uses for Emory Healthcare Network, the largest health system in Georgia with seven hospitals responsible for nearly 5 million patient encounters in 2017. Emory Health Network and Emory LITS have had an established process for judicious sharing of protected electronic patient health data that was leveraged for the development of this HIV disease registry.

Therefore, the issue with lack of resources was best resolved by introducing experienced and knowledgeable personnel that was not available through the healthcare system.

Limiting Strain on the Healthcare Business

Secondly, the case study revealed the disease registry was built in such a way that it does not impact or place a strain on the health care system's enterprise. This was accomplished by setting the disease registry on its own server, but having the server remain within the healthcare system's infrastructure. A data staging area owned by the healthcare system was utilized for data to be copied and transferred from the healthcare system's data sources. In this way, the data used in healthcare could not be altered by using and sharing data through the HIV disease registry. Having this separation resolved the potential impact or strain on the data used every day in healthcare management. Additionally, maintaining the physical hardware on the healthcare system's enterprise meant that data would be contained within the organization's infrastructure while stored on the disease registry's server. This arrangement can be used as a model for how other disease registries can store its data while maintaining linkage to the healthcare system's EHR.

Providing Continuity and Management

Thirdly, stability of Emory CFAR provides the continuity and funding to support the HIV disease registry operations. The organization receives nearly $4 million in funding from NIH that it uses to support and facilitate research in HIV/AIDS. The mission statement includes the support for this HIV disease registry that acknowledges it plays a pivotal part in meeting the aims of Emory CFAR to facilitate secondary analysis research. In 2013, the NIH awarded nearly $200,000 to Emory CFAR for the development and implementation of its clinical disease registry. The cost of developing the disease registry along with the annual cost of operations of $152,000 is remarkably low for the scope of this type of project further demonstrating that a successful HIV disease registry can be built and managed without significant investments.

Assuring Accountability for Data Uses

Fourthly, this case study examined how Emory CFAR developed an accountability process to meet the requirements of the healthcare system. These processes could be replicated for new disease registries. A research advisory council was assembled with patient advocates consisting of doctors and healthcare leaders to review the scientific merit and propriety of all data requests. This safeguard was to promote practical and beneficial secondary uses of data that would contribute scientific merit, improvements to health outcomes, and process improvements for the delivery of care. Another responsibility of the council is to prevent needless, unwarranted, or duplicate applications for secondary uses of patient health data. In addition, a formal data request and scope of work forms were created to track all uses of data. Being able to trace all data to users was important for the buy-in of the healthcare system. Developing a disease registry would benefit from an oversight council and with data accountability documents such as a data request form and scope of work document.

Administering Data Governance

Fifthly, this case study describes the governance system enacted by Emory CFAR that has been endorsed by the data stewards of the healthcare system. The establishment of a data management agreement in an MOU forged a legal partnership between the two parties that allowed Emory CFAR to distribute data on behalf of the healthcare system. Furthermore, end users of data entered into DUAs that would limit and control the uses of data. This allowed Emory CFAR to maintain regulation and HIPAA-compliance of secondary analyses of data for public health research. The DUAs also incorporated a data destruction statement to secure the elimination of data at the conclusion of the research. These legal documents contributed to the endorsement of the HIV disease registry.

Codifying this experience and the governance processes of sharing data with external partners would be a valuable tool for public health research in a similar environment as described in this case study.

Securing Data

Lastly, this case study explains the technical, physical and administrative security requirements needed to protect patient health data. Technical and physical security for the HIV disease registry were bolstered by keeping the data within the healthcare organization's informatics enterprise. With this approach, the data stewards at the healthcare organization are able to maintain some control of securing the data. Administrative security was implemented for the HIV disease registry database, but users are also required to gain user authorization and authentication through the healthcare organization to be able to access the database that exists on the secured informatics enterprise of the hospital.

7.3 Limitations

This case study on the Emory CFAR HIV disease registry is not a universal solution. The case itself was a narrow focus design that offered a solution that was customized for these two organizations to improve opportunities in public health research for HIV/AIDS at this healthcare system. Many of the components of the disease registry attend specifically to the uniqueness of each of these two institutions. Therefore, this informatics solution may not precisely translate to other healthcare organizations. Not all healthcare systems use the same information management systems as the healthcare system described in this case study and thus this case study cannot accommodate the myriad of possible combinations of an informatics enterprise. In addition, the targeted data variables vary for each disease. There are different standard guidelines for care for each disease and condition using their own case definitions and healthcare metrics. To overcome these limitations, data and technical strategies should be customized to meet the needs of different diseases and conditions with particular attention to the informatics ecosystem and data sources that exist in order to effectively use patient health data in secondary analyses. The HIV disease registry selected many types of labs to include in its database after consulting with subject matter experts to determine the needs for current and potential future research. Involving subject matter experts would make it simple to select specific labs and procedures to target the evaluation of a specific disease or condition. Selecting all clinical labs that exist in an EHR is also an easy solution if the capacity and resources exist for a disease registry. However, because data should be evaluated before entering a production database used in research, the more data there are the more resources and time invested in evaluating the quality of data.

Working with two institutions that have different organizational policies and processes is another limitation of this case study. Differing policies introduce a complexity in information accessibility within an inter-institutional relationship. The ambiguity of data ownership is an aspect of this limitation. In this case study, the healthcare system owns the data contained in the data sources while the data user owns data produced through research resulting from data analyses. However, there is uncertainty about the ownership of data that resides in staging areas and the disease registry itself. It was unclear who the data stewards are for all the spaces that data were at rest or in transit. Establishing a data use agreement that addresses these data ownership issues can help overcome this limitation. Another aspect of this limitation is that the benefit to the healthcare system may be hindered because secondary uses of data for research are not recorded in the EHR or other data source, so public health knowledge gained through secondary analyses is not shared at the point of care delivery. This is a business decision that would have to be made by the healthcare organization for whether the results of data analyses should be included in the EHR so that knowledge from analyses can be more easily accessed by front-line providers.

Access to the data introduced technical limitations. Reviewing data required installing a front end user interface as well as knowledge of SQL programming. Installing the Toad application that served as the front end user interface was difficult and required changing SQL code in its program library used in the program installation. To query data in the HIV disease registry also required generating SQL scripts to link relational data tables within the disease registry. One solution to overcoming this limitation is improvising and installing a front-end application that simplifies access and visualization of data contained in the disease registry. A front-end application such as Tableau or R can be linked to the disease registry so that data can be more easily accessed for public health research without SQL querying expertise.

7.4 Recommendations

There were several things to note in examining the BAA for this case study. The BAA does not define who gives approval for the release of data to Emory CFAR or to end-users for secondary uses in public health research. This should be included in negotiating the MOU between Emory CFAR and the healthcare system. The process at Emory University for approvals of PHI disclosure is typically a combination of Emory IRB, the healthcare system's research oversight committee, and the healthcare system's office of compliance. The current process for data sharing involves the healthcare system departments of quality and information services approving specific dataset to be transferred to the HIV disease registry, however, these departments are not involved in permission for uses. This should be more clearly defined in the working MOU.

The HIV disease registry may benefit from further evaluation after completion of its planned future enhancements. The scope of this case study did not include a comprehensive assessment of interoperability of systems, standardization of data and systems, and the scalability of the disease registry to evolve with public health. The HIV disease registry has interoperability and standardization processes in development that are not complete and an evaluation would be premature. The registry currently uses Logical Observation Identifiers Names and Codes (LOINC) and Systematized Nomenclature of Medicine (SNOMED) standards, but is also embarking on the integration of the Observational Medical Outcomes Partnership (OMOP) to harmonize the disparate systems to a standardized data model and vocabulary. The OMOP solution would expand the HIV disease registry for scalability to incorporate new observational systems or connect the disease registry with national registries. The primary intent for OMOP adoption is to integrate HIV data from two other healthcare systems to create a singular disease registry about PLWH from across the three healthcare organizations. A future case study can evaluate this standardization process so that this model could be shared as a template for other disease registries to follow a similar transition.

A further opportunity for the HIV disease registry and a future case study is the incorporation of free text captured by the healthcare system that comprises nearly 80% of all health data in the EHR. Public health informaticians can develop tools and mechanics for natural language processing to realize unstructured data in a discrete and searchable format. Data elements could be parsed from unstructured data in EHRs to create discrete variables that encompass the important information within narrative text. Alternatively, unmodified and unstructured data can be stretched across clusters of parallel systems to enable rapid querying of huge volumes of information without having to enforce structure to narrative text.

Developing a disease registry similar to this HIV disease registry for other chronic diseases and conditions is an important opportunity for non-HIV investigators. This case study provides a useful template despite targeting only HIV/AIDS because differences in data requirements for chronic diseases are generally marginal from a data collection perspective. Case definitions and risks for diseases are focused on clinical characteristics for which standardized guidelines for care are developed and are being captured in the EHR for routine care management of chronic diseases. Disease registries linked to EHRs would include these clinical characteristics that are contained within lab results, diagnoses, problem lists, patient encounters, treatments, and procedures. These are the necessary data domains for understanding disease pathogenesis and progression as well as impact of therapeutic and behavioral interventions so a disease registry can use this case study to establish a database for any chronic disease.

Disease registries can play an important transformative role in any healthcare environment to support clinical, translational and pharmaceutical research that can improve health outcomes. These types of databases are efficient informatics tools for healthcare epidemiology that involve surveillance, prevention, and control of adverse events in healthcare to study causes, factors and outcomes of healthcare and healthcare delivery. Clinical, administrative and financial information are valuable data contained in EHRs that can be loaded into a linked disease registry to expand capabilities for big data research initiatives.

7.5 Conclusion

The Emory CFAR HIV disease registry exemplifies a successful model for data sharing that navigated the legal landscape of secondary uses of protected electronic patient health data. This solution was able to overcome a defensive healthcare system that is cautious of releasing data and control of data accountability and governance to an external partner. The HIV disease registry team developed key strategies for data governance, accountability and security that provided assurances to the data stewards to enable their buy-in to share PHI. Throughout the process of developing the Emory CFAR disease registry, its project team experienced nearly parallel challenges as seen with the eID clinical data warehouse in Chicago. The regulatory and political barrier is a continuous process that evolves and adapts as the data stewards, data, and data sources change. The HIV disease registry team will always be chasing these issues because they are not involved in the change management communications in the healthcare system. There were also barriers presented by the healthcare system's lack of resources, experience and knowledge to support secondary uses of patient health data. The project also faced challenges of working with a proprietary EHR system that would not share its data model, data dictionary or documentation. Using the three key strategies, the HIV disease registry team was successful in resolving and circumventing these barriers to create a replicable model for how data can be shared across organizations.

The resulting database was capable of secure cross-institutional sharing of protected health information to provide scientifically robust data to public health investigators and healthcare leadership. Additionally, the database was built on a framework that can be translated for use with other diseases. This case study provides guidance to other academic researchers for establishing EHR-linked disease registries with healthcare partners. The disease registry is designed by employing data and technical standards (LOINC, SNOMED, and soon to be OMOP) that will allow it to connect to other data sources for uni- or bi-directional exchange of data. This exchange of data is accomplished with an interoperable framework to facilitate data transfer between the enterprises as well as other public health systems such as regional health information organizations or national surveillance databases. Successful negotiations with data stewards to obtain access to data are accomplished by engaging data stewards with strategies for interoperability, governance, accountability, and security.

1. Malmberg, E.D., et al., *Improving HIV/AIDS Knowledge Management Using EHRs.* Online J Public Health Inform, 2012. **4**(3).

2. Muther, J., *2012 Annual Ryan White HIV/AIDS Program Data Report for Grady Health System Ponce Infectious Disease Program*, H.A.P. U.S. Dept. of Health and Human Services, Ryan White Program, Editor. 2013.

3. Hanna, K.A., S.; Davidson-Maddox, S., *Using EMRs to Bridge Patient Care and Research*. 2005, FasterCures.

4. PriceWaterhouseCoopers, *Transforming Healthcare Through Secondary Use of Health Data*, in *Health Industries*. 2009.

5. U.S. Centers for Disease Control and Prevention. *HIV/AIDS Basic Statistics*. 2018 [cited 2018; Available from: https://www.cdc.gov/hiv/basics/statistics.html.

6. U.S. Centers for Disease Control and Prevention, *HIV in the United States: The Stages of Care*, H.A.P. U.S. Dept. of Health and Human Services, Editor. July, 2012.

7. Ritchwood, T.D., K.G. Bishu, and L.E. Egede, *Trends in healthcare expenditure among people living with HIV/AIDS in the United States: evidence from 10 Years of nationally representative data.* Int J Equity Health, 2017. **16**(1): p. 188.

8. Farnham, P.G., et al., *Lifetime costs and quality-adjusted life years saved from HIV prevention in the test and treat era.* J Acquir Immune Defic Syndr, 2013. **64**(2): p. e15-8.

9. Nakagawa, F., et al., *Projected Lifetime Healthcare Costs Associated with HIV Infection.* PLoS One, 2015. **10**(4): p. e0125018.

10. Dombrowski, J.C., et al., *Population-based metrics for the timing of HIV diagnosis, engagement in HIV care, and virologic suppression.* AIDS, 2012. **26**(1): p. 77-86.

11. Hall, H.I., et al., *HIV care visits and time to viral suppression, 19 U.S. jurisdictions, and implications for treatment, prevention and the national HIV/AIDS strategy.* PLoS One, 2013. **8**(12): p. e84318.

12. Hall, H.I., et al., *Differences in human immunodeficiency virus care and treatment among subpopulations in the United States.* JAMA Intern Med, 2013. **173**(14): p. 1337-44.

13. Thompson, M., *Phase I Progress Report: Building the Strategy to End AIDS in Fulton County*. 2015, Fulton County Task Force on HIV/AIDS.

14. Wortley, P.C., J.; Harris, B.; Lovell, K.; Lyons, W.; Ealey, J.; Cruz-Lopez, M.; Ramos, M.; Mootry, B.; Momah, V.; Seabolt, M.; McKinley-Beach, L.; Anderson, D.C.; Arromand, A.A.; Vollman, J.; Ervin, C.; Whyte, K., *Georgia Integrated HIV Prevention & Care Plan 2017-2021*. 2016, Georgia Dept. of Public Health HIV Surveillance Section.

15. Baseman, J., Revere, D., Painter, I., *Big Data in the Era of Health Information Exchanges: Challenges and Opportunities for Public Health.* Informatics, 2017. **4**(39).

16. Mallow, G.K., *The value of healthcare data for secondary uses in clinical research and development*, in *HIMSS 2012 Annual Meeting.* 2012.

17. Jamoom E, Y.N., *Table of Electronic Health Record Adoption and Use Among Office-based Physicians in the U.S., by State: 2015 National Electronic Health Records Survey.* 2016.

18. Hillestad, R., et al., *Can electronic medical record systems transform health care? Potential health benefits, savings, and costs.* Health Aff (Millwood), 2005. **24**(5): p. 1103-17.

19. Bauer-Mehren, A., et al., *Network analysis of unstructured EHR data for clinical research.* AMIA Jt Summits Transl Sci Proc, 2013. **2013**: p. 14-8.

20. Murphy, E.F.I., FI; O'Donnell, WR, *An Electronic Medical Records System for Clinical Research and the EMR–EDC Interface.* Investigative Ophthalmology & Visual Science, 2007. **48**(10): p. 4383-4389.

21. van Panhuis, W.G., et al., *A systematic review of barriers to data sharing in public health.* BMC Public Health, 2014. **14**: p. 1144.

22. Garg, A.X., et al., *Effects of computerized clinical decision support systems on practitioner performance and patient outcomes: a systematic review.* JAMA, 2005. **293**(10): p. 1223-38.

23. Humphries, K.H., et al., *Co-morbidity data in outcomes research: are clinical data derived from administrative databases a reliable alternative to chart review?* J Clin Epidemiol, 2000. **53**(4): p. 343-9.

24. Kiragga, A.N., et al., *Quality of data collection in a large HIV observational clinic database in sub-Saharan Africa: implications for clinical research and audit of care.* J Int AIDS Soc, 2011. **14**: p. 3.

25. Logan, J.R. and D.A. Lieberman, *The use of databases and registries to enhance colonoscopy quality.* Gastrointest Endosc Clin N Am, 2010. **20**(4): p. 717-34.

26. Brown, P.J.B. and P. Sonksen, *Evaluation of the quality of information retrieval of clinical findings from a computerized patient database using a semantic terminological model.* Journal of the American Medical Informatics Association, 2000. **7**(4): p. 392-403.

27. Chaudhry, B., et al., *Systematic review: impact of health information technology on quality, efficiency, and costs of medical care.* Ann Intern Med, 2006. **144**(10): p. 742-52.

28. Hayrinen, K., K. Saranto, and P. Nykanen, *Definition, structure, content, use and impacts of electronic health records: a review of the research literature.* Int J Med Inform, 2008. **77**(5): p. 291-304.

29. Leitheiser, R., *Data Quality in Health Care Data Warehouse Environments*, in *34th Hawaii International Converence on System Sciences*. 2001: Hawaii, USA.

30. Willig, J.H., *Data at the HIV-Research and Informatics Service Center (RISC)*. 2011.

31. Ancker, J.S., et al., *Root causes underlying challenges to secondary use of data.* AMIA Annu Symp Proc, 2011. **2011**: p. 57-62.

32. El Fadly, A., et al., *Integrating clinical research with the Healthcare Enterprise: from the RE-USE project to the EHR4CR platform.* J Biomed Inform, 2011. **44 Suppl 1**: p. S94-102.

33. Kopcke, F., et al., *Evaluation of data completeness in the electronic health record for the purpose of patient recruitment into clinical trials: a retrospective analysis of element presence.* BMC Med Inform Decis Mak, 2013. **13**: p. 37.

34. Murphy, E.C., F.L. Ferris, 3rd, and W.R. O'Donnell, *An electronic medical records system for clinical research and the EMR EDC interface.* Invest Ophthalmol Vis Sci, 2007. **48**(10): p. 4383-9.

35. Ebidia, A., et al., *Getting data out of the electronic patient record: critical steps in building a data warehouse for decision support.* Proc AMIA Symp, 1999: p. 745-9.

36. Deloitte, *Secondary uses of Electronic Health Record (EHR) data in Life Sciences*. 2009.

37. Hersh, W., *Electronic health records facilitate development of disease registries and more.* Clin J Am Soc Nephrol, 2011. **6**(1): p. 5-6.

38. Navaneethan, S.D., et al., *Development and validation of an electronic health record-based chronic kidney disease registry.* Clin J Am Soc Nephrol, 2011. **6**(1): p. 40-9.

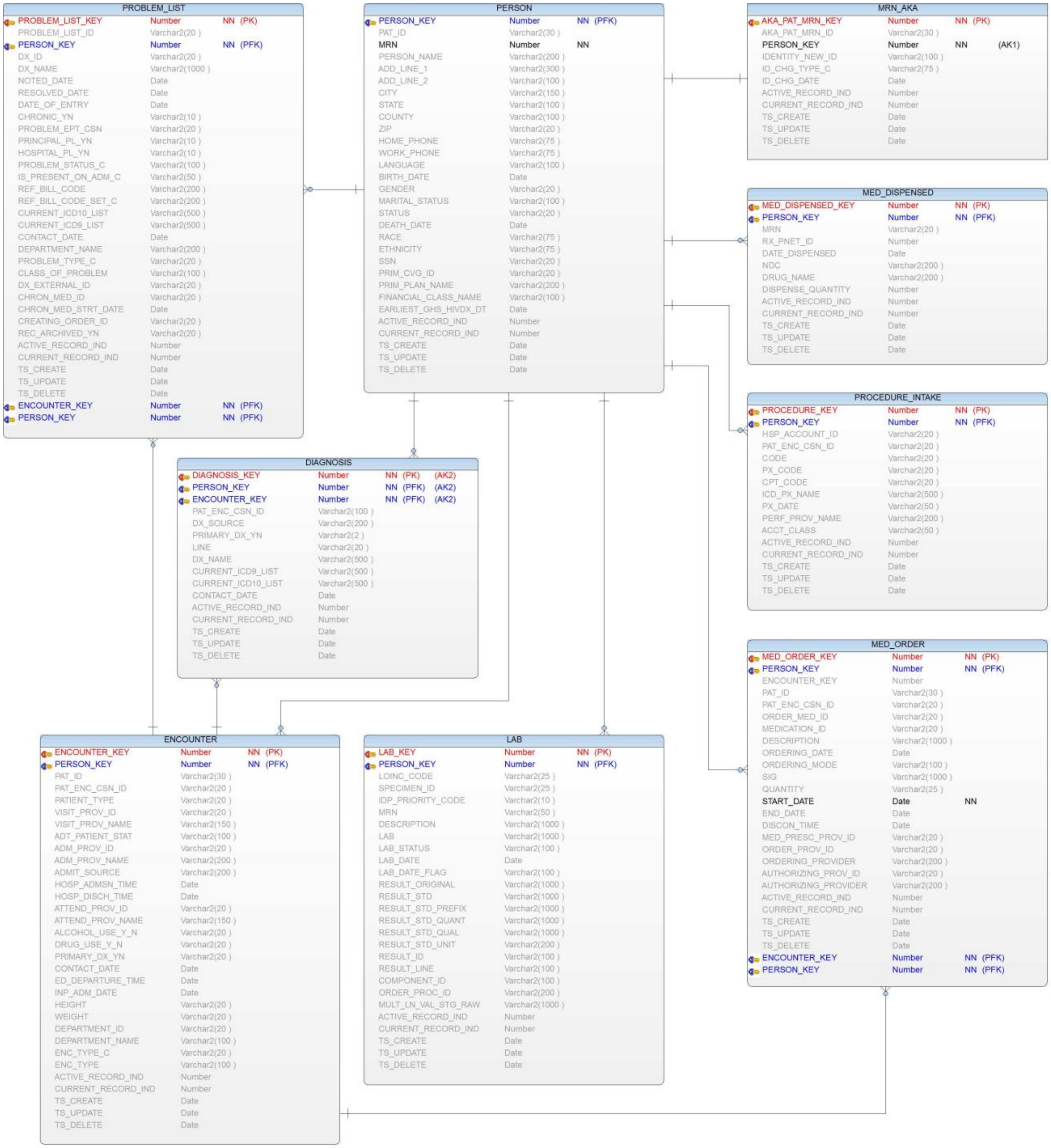
39. Wisniewski, M.F., et al., *Development of a clinical data warehouse for hospital infection control.* J Am Med Inform Assoc, 2003. **10**(5): p. 454-62.

40. Zozus, M.N.H., W.E; Green, B.B; Kahn, M.G.; Richesson, R.L.; Rusinocovitch, S.A.; Simon, G.E.; Smerek, M.M., *Assessing Data Quality for Healthcare Systems Data Used in Clinical Research (Version 1.0)*. 2014, NIH Health Care Systems Research Collaboratory.

8 APPENDIX I: INTERVIEW QUESTIONS

1. Principal investigator
   a. Describe the relationship and limitations between the two organizations.
   b. Describe the development of the disease registry.
   c. How is the disease registry managed? By who?
   d. How are data from the disease registry used to support public health?
   e. How are data uses governed?
   f. What is the business process for data extraction?
   g. Who owns the data in the disease registry?
   h. Who owns the data generated by research using patient health data?
   i. What are the most significant barriers/challenges experienced in the development and management of the disease registry?
2. Informatics architect
   a. Describe the database schema and infrastructure.
   b. What are the key advantages/disadvantages of the framework you selected?
   c. Describe the technical environment.
   d. What are the advantages/disadvantages to the ecosocial and technical environments?
   e. Describe the ETL process(es) for transferring patient health data to the disease registry.
   f. What is the current technical strategy of the disease registry? What are the plans for the future?
   g. What are the challenges and limitations of the technical infrastructure of the disease registry?
   h. How does the disease registry comply with HIPAA Security and Safety requirements?
3. Data analyst
   a. What are the data sources?
   b. Describe the relational data tables.
   c. How are data transferred and shared?
   d. How are data stored and managed?
   e. What are the security measures in place for the disease registry?
   f. Describe the technical environment.
   g. What is the technical process for data extraction?
   h. How do you ensure the quality of data to be used in research?
4. Clinical SME
   a. How were the data selected to be included in the HIV disease registry?
   b. What are the limitations of the selected data elements for use in public health?
   c. What is the case definition used for the inclusion of HIV patients in the disease registry?
   d. How is your role important?
   e. How many projects have been supported by the disease registry? Describe some of the key projects.

9 Appendix II: Physical Data Model



PROBLEM_LIST

PROBLEM_LIST_KEY	Number	NN	(PK)
PROBLEM_LIST_ID	Varchar2(20)		
PERSON_KEY	Number	NN	(PFK)
DX_ID	Varchar2(20)		
DX_NAME	Varchar2(1000)		
NOTED_DATE	Date		
RESOLVED_DATE	Date		
DATE_OF_ENTRY	Date		
CHRONIC_YN	Varchar2(10)		
PROBLEM_EPT_CSN	Varchar2(20)		
PRINCIPAL_PL_YN	Varchar2(10)		
HOSPITAL_PL_YN	Varchar2(10)		
PROBLEM_STATUS_C	Varchar2(100)		
IS_PRESENT_ON_ADM_C	Varchar2(50)		
REF_BILL_CODE	Varchar2(200)		
REF_BILL_CODE_SET_C	Varchar2(200)		
CURRENT_ICD10_LIST	Varchar2(500)		
CURRENT_ICD9_LIST	Varchar2(500)		
CONTACT_DATE	Date		
DEPARTMENT_NAME	Varchar2(200)		
PROBLEM_TYPE_C	Varchar2(20)		
CLASS_OF_PROBLEM	Varchar2(100)		
DX_EXTERNAL_ID	Varchar2(20)		
CHRON_MED_ID	Varchar2(20)		
CHRON_MED_STRT_DATE	Date		
CREATING_ORDER_ID	Varchar2(20)		
REC_ARCHIVED_YN	Varchar2(20)		
ACTIVE_RECORD_IND	Number		
CURRENT_RECORD_IND	Number		
TS_CREATE	Date		
TS_UPDATE	Date		
TS_DELETE	Date		
ENCOUNTER_KEY	Number	NN	(PFK)
PERSON_KEY	Number	NN	(PFK)

PERSON

PERSON_KEY	Number	NN	(PFK)
PAT_ID	Varchar2(30)		
MRN	Number	NN	
PERSON_NAME	Varchar2(200)		
ADD_LINE_1	Varchar2(300)		
ADD_LINE_2	Varchar2(100)		
CITY	Varchar2(150)		
STATE	Varchar2(100)		
COUNTY	Varchar2(100)		
ZIP	Varchar2(20)		
HOME_PHONE	Varchar2(75)		
WORK_PHONE	Varchar2(75)		
LANGUAGE	Varchar2(100)		
BIRTH_DATE	Date		
GENDER	Varchar2(20)		
MARITAL_STATUS	Varchar2(100)		
STATUS	Varchar2(20)		
DEATH_DATE	Date		
RACE	Varchar2(75)		
ETHNICITY	Varchar2(75)		
SSN	Varchar2(20)		
PRIM_CVG_ID	Varchar2(20)		
PRIM_PLAN_NAME	Varchar2(200)		
FINANCIAL_CLASS_NAME	Varchar2(100)		
EARLIEST_GHS_HIVDX_DT	Date		
ACTIVE_RECORD_IND	Number		
CURRENT_RECORD_IND	Number		
TS_CREATE	Date		
TS_UPDATE	Date		
TS_DELETE	Date		

MRN_AKA

AKA_PAT_MRN_KEY	Number	NN	(PK)	
AKA_PAT_MRN_ID	Varchar2(30)			
PERSON_KEY	Number	NN		(AK1)
IDENTITY_NEW_ID	Varchar2(100)			
ID_CHG_TYPE_C	Varchar2(75)			
ID_CHG_DATE	Date			
ACTIVE_RECORD_IND	Number			
CURRENT_RECORD_IND	Number			
TS_CREATE	Date			
TS_UPDATE	Date			
TS_DELETE	Date			

MED_DISPENSED

MED_DISPENSED_KEY	Number	NN	(PK)
PERSON_KEY	Number	NN	(PFK)
MRN	Varchar2(20)		
RX_PNET_ID	Number		
DATE_DISPENSED	Date		
NDC	Varchar2(200)		
DRUG_NAME	Varchar2(200)		
DISPENSE_QUANTITY	Number		
ACTIVE_RECORD_IND	Number		
CURRENT_RECORD_IND	Number		
TS_CREATE	Date		
TS_UPDATE	Date		
TS_DELETE	Date		

PROCEDURE_INTAKE

PROCEDURE_KEY	Number	NN	(PK)
PERSON_KEY	Number	NN	(PFK)
HSP_ACCOUNT_ID	Varchar2(20)		
PAT_ENC_CSN_ID	Varchar2(20)		
CODE	Varchar2(20)		
PX_CODE	Varchar2(20)		
CPT_CODE	Varchar2(20)		
ICD_PX_NAME	Varchar2(500)		
PX_DATE	Varchar2(50)		
PERF_PROV_NAME	Varchar2(200)		
ACCT_CLASS	Varchar2(50)		
ACTIVE_RECORD_IND	Number		
CURRENT_RECORD_IND	Number		
TS_CREATE	Date		
TS_UPDATE	Date		
TS_DELETE	Date		

DIAGNOSIS

DIAGNOSIS_KEY	Number	NN	(PK)	(AK2)
PERSON_KEY	Number	NN	(PFK)	(AK2)
ENCOUNTER_KEY	Number	NN	(PFK)	(AK2)
PAT_ENC_CSN_ID	Varchar2(100)			
DX_SOURCE	Varchar2(200)			
PRIMARY_DX_YN	Varchar2(2)			
LINE	Varchar2(20)			
DX_NAME	Varchar2(500)			
CURRENT_ICD9_LIST	Varchar2(500)			
CURRENT_ICD10_LIST	Varchar2(500)			
CONTACT_DATE	Date			
ACTIVE_RECORD_IND	Number			
CURRENT_RECORD_IND	Number			
TS_CREATE	Date			
TS_UPDATE	Date			
TS_DELETE	Date			

MED_ORDER

MED_ORDER_KEY	Number	NN	(PK)
PERSON_KEY	Number	NN	(PFK)
ENCOUNTER_KEY	Number		
PAT_ID	Varchar2(30)		
PAT_ENC_CSN_ID	Varchar2(20)		
ORDER_MED_ID	Varchar2(20)		
MEDICATION_ID	Varchar2(20)		
DESCRIPTION	Varchar2(1000)		
ORDERING_DATE	Date		
ORDERING_MODE	Varchar2(100)		
SIG	Varchar2(1000)		
QUANTITY	Varchar2(25)		
START_DATE	Date	NN	
END_DATE	Date		
DISCON_TIME	Date		
MED_PRESC_PROV_ID	Varchar2(20)		
ORDER_PROV_ID	Varchar2(20)		
ORDERING_PROVIDER	Varchar2(200)		
AUTHORIZING_PROV_ID	Varchar2(20)		
AUTHORIZING_PROVIDER	Varchar2(200)		
ACTIVE_RECORD_IND	Number		
CURRENT_RECORD_IND	Number		
TS_CREATE	Date		
TS_UPDATE	Date		
TS_DELETE	Date		
ENCOUNTER_KEY	Number	NN	(PFK)
PERSON_KEY	Number	NN	(PFK)

ENCOUNTER

ENCOUNTER_KEY	Number	NN	(PK)
PERSON_KEY	Number	NN	(PFK)
PAT_ID	Varchar2(30)		
PAT_ENC_CSN_ID	Varchar2(20)		
PATIENT_TYPE	Varchar2(20)		
VISIT_PROV_ID	Varchar2(20)		
VISIT_PROV_NAME	Varchar2(150)		
ADT_PATIENT_STAT	Varchar2(100)		
ADM_PROV_ID	Varchar2(20)		
ADM_PROV_NAME	Varchar2(200)		
ADMIT_SOURCE	Varchar2(200)		
HOSP_ADMSN_TIME	Date		
HOSP_DISCH_TIME	Date		
ATTEND_PROV_ID	Varchar2(20)		
ATTEND_PROV_NAME	Varchar2(150)		
ALCOHOL_USE_Y_N	Varchar2(20)		
DRUG_USE_Y_N	Varchar2(20)		
PRIMARY_DX_YN	Varchar2(20)		
CONTACT_DATE	Date		
ED_DEPARTURE_TIME	Date		
INP_ADM_DATE	Date		
HEIGHT	Varchar2(20)		
WEIGHT	Varchar2(20)		
DEPARTMENT_ID	Varchar2(20)		
DEPARTMENT_NAME	Varchar2(100)		
ENC_TYPE_C	Varchar2(20)		
ENC_TYPE	Varchar2(100)		
ACTIVE_RECORD_IND	Number		
CURRENT_RECORD_IND	Number		
TS_CREATE	Date		
TS_UPDATE	Date		
TS_DELETE	Date		

LAB

LAB_KEY	Number	NN	(PK)
PERSON_KEY	Number	NN	(PFK)
LOINC_CODE	Varchar2(25)		
SPECIMEN_ID	Varchar2(25)		
IDP_PRIORITY_CODE	Varchar2(10)		
MRN	Varchar2(50)		
DESCRIPTION	Varchar2(1000)		
LAB	Varchar2(1000)		
LAB_STATUS	Varchar2(100)		
LAB_DATE	Date		
LAB_DATE_FLAG	Varchar2(100)		
RESULT_ORIGINAL	Varchar2(1000)		
RESULT_STD	Varchar2(1000)		
RESULT_STD_PREFIX	Varchar2(1000)		
RESULT_STD_QUANT	Varchar2(1000)		
RESULT_STD_QUAL	Varchar2(1000)		
RESULT_STD_UNIT	Varchar2(200)		
RESULT_ID	Varchar2(100)		
RESULT_LINE	Varchar2(100)		
COMPONENT_ID	Varchar2(100)		
ORDER_PROC_ID	Varchar2(200)		
MULT_LN_VAL_STG_RAW	Varchar2(1000)		
ACTIVE_RECORD_IND	Number		
CURRENT_RECORD_IND	Number		
TS_CREATE	Date		
TS_UPDATE	Date		
TS_DELETE	Date		

10 APPENDIX III: STAKEHOLDER PROFILES

Stakeholder Category	Phase with Maximum Impact	Role / how project affects Group	Concerns or Issues	Stakeholder Strategy
Data Owners	Development	• Authorizes use of data • Will manage the data exchange	• Patient protection • System security • Data safety • Maintenance costs • Data use agreement • Data security/integrity • Maintenance of local systems and equipment	Project Team, Medical Directors, and Managing Sponsors will present proposal and address their concerns; and Project Team will work to establish precedents of data use.
Funding Sponsors	Discovery	• Authorizes project • Provides initial and ongoing funding	• Utility of the system • Cost • Feasibility/ interoperability • Schedule/Timeline	Quarterly reporting; or as determined by sponsor.
Managing Sponsors	All	• System owners • Primary end users • Manages project • Negotiates funding and data use	• Utility of the system • Feasibility/interoperability • Gaining adequate funding • Cost • Achieving buy-in • Data use agreement • Timeline; achieving milestones • Patient protection, HIPAA • Data Quality • System Security	Inform at routine bi-weekly meetings, Project Team aids in setting up meetings for key relationships, Project Team provides documents/reports/budgets that they will use in negotiations.
Subject Matter Experts	Various	• Provides expertise	• Various, depending on expertise	Inform at routine monthly meetings.
Project Team	All	• Project management • Manage funds	• Feasibility/ interoperability • Usefulness of design • Developing and maintaining key relationships	• Monitor and inform of progress at regularly schedule meetings

Stakeholder Category	Phase with Maximum Impact	Role / how project affects Group	Concerns or Issues	Stakeholder Strategy
			• Timeline; achieving milestones • Gaining adequate funding • Data use agreement • System security • Data quality	• Ensure communication delivery • Organize meetings & schedules • Coordinate high level interaction with the Project team, Sponsors and Subject Matter Experts
Informatics Team	All	• Systems design	• Systems budget • System security • Utility • Feasibility/interoperability • Achieving buy-in • Data use agreements in place • Data quality	Use informaticians to aid in presentations with sponsors and key relationships.
Regulatory Agencies	Implementation	• Determines compliance for regulations and patient protection	• Regulatory • Patient protection, HIPAA • System security	Ongoing communications for regulatory management; annual regulatory renewals; communicate changes in the project as appropriate.

11 Appendix IV: Project Plan

11.1 Work Breakdown Structure

Database Design	Database Implementation	ETL Process Design & Implementation	Post Production
TASKS	TASKS	TASKS	TASKS
Start Up	Create a Db instance	Create Lab ETL (Ultra C)	Gather user interface requirements
Data negotiation & agreement	Create logical & physical data models	Create Demographics ETL	Create graphic user interface
Gather data requirements & preliminary data	Normalize data	Create Visit & Consult ETL	Gather CNICS requirements
Establish business processes & policies	Initial Db launch	Create Conditions & Problem List ETL	Establish process for CNICS connection
Create data dictionary	Develop a data utility script	Create Medication ETL	Develop end user documentation
Data cleaning	Execute sample data production	Create Procedures, Screenings, & Immunizations ETL	Train end users
System servicing	Migrate data to healthcare organization's servers	Create Lab ETL (Beaker)	Assess risk management & mitigation strategy
Data monitoring	Design reporting & parsing functionality	Create Hospitalizations ETL	Establish quality check schedule & policies
Data provision	Definitive Db launch	Prepare Db configuration information	Establish quality reporting process
Verify HIPAA compliance	System testing & validation	Prepare data sources & connections	Establish & maintain security policies and SOPs
Establish data use agreement for end users	Construct data report templates	Prepare ETL catalog, fact tables, & dimension tables	Establish IT security reporting process
Establish process for end users to access data	Train database support staff	Deploy & test ETL jobs	

11.2 Work breakdown structure dictionary

TASK	ASSIGNED TO
DATABASE DESIGN	
Project startup	Project Team, Informatics Team
Negotiate terms for the use of healthcare data and establish a data use agreement.	PI, Architect, Project Manager
Gather the data requirements from sources from SMEs and IDP investigators for the construction of the data model; construct preliminary data.	Business Analyst, Informatics Team, Project Manager
Distinguish the high level business processes and policies.	Architect, Project Manager
Create a data dictionary that will serve to normalize data and support future interoperability.	Informatics Team
Data cleaning: Weekly, imported data will be reviewed for systematic and random errors and corrective action will be undertaken.	Data Manager
Servicing: Any system malfunctions, security breaches, software upgrades or other maintenance issues will be managed accordingly.	Data Manager
Data monitoring: the database will be reviewed quarterly for quality assurance and validity.	PI, Architect, Clinical Informatics Consultant, Data Manager, Database Administrator
Data provision: Upon approval of data requests from the principal investigator of the QA/PI study, data elements from the IDP Db will be provided in the required format directly to the investigator for analysis.	Data Manager, Data Analyst
Validate and test the system for HIPAA compliance	Architect, Data Manager
Establish a data use agreement for end users.	PI, Project Manager
Establish a process for end users to gain access to data.	PI, Project Manager

TASK	ASSIGNED TO
DATABASE IMPLEMENTATION	
Using the existing database instance, the initial HIV disease registry instance will be developed.	Architect, Business Analyst, EPIC Consultant
Create contextual and logical data models.	Architect, Project Manager
Normalization: Data will be harmonized from disparate sources so that it is represented in the same form.	Architect, Data Manager
Initial database test bed deployed off site.	Architect, Programmer
Develop a data utility script to parse test source files and populate the test database.	Architect, Programmer
Sample data production will be executed to test the capacity and workflow of the database.	Architect, Programmer
The database will be migrated to the servers.	Architect, Programmer
Reports will be produced and stored to be parsed into a hierarchy of components in order to populate the database.	Architect, Programmer
Definitive database launch.	Architect, Programmer, Data Manager
System testing and validation.	Architect, Programmer, Data Manager
A set of custom reporting templates will be modeled using the existing templates from the Diabetes Db.	Architect, Programmer, Data Manager
Train the database support staff.	Architect, Programmer, Data Manager

TASK	ASSIGNED TO
ETL PROCESS DESIGN & IMPLEMENTATION	
Create Laboratory ETL process (Ultra C).	Architect, Programmer, Data Manager
Create Demographics ETL process.	Architect, Programmer, Data Manager
Create Visits and Consultations ETL processes.	Architect, Programmer, Data Manager
Create Conditions and Problem List ETL processes.	Architect, Programmer, Data Manager
Create Medication ETL process.	Architect, Programmer, Data Manager
Create Procedures, Screenings, and Immunizations ETL processes.	Architect, Programmer, Data Manager
Create Laboratory ETL process (Beaker).	Architect, Programmer, Data Manager
Create Hospitalizations ETL process.	Architect, Programmer, Data Manager
These apply to all the above ETL processes:	
Prepare database configuration information.	Architect, Programmer, Data Manager
Prepare data sources and connections.	Architect, Programmer, Data Manager
Prepare ETL catalog, fact tables, and dimension tables.	Architect, Programmer, Data Manager
Deploy and test ETL jobs.	Architect, Programmer, Data Manager

TASK	ASSIGNED TO
POST PRODUCTION	
Gather the user requirements to develop the graphic user interface.	Architect, Programmer, Project Manager
Create a graphic user interface.	Architect, Programmer
Gather the data and connectivity requirements for linking to CNICS.	Architect, Programmer, Project Manager
Establish the processes for CNICS connection.	Architect, Programmer
Develop end user documentation.	Architect, Project Manager
Train end users.	Architect, Programmer
Assess risk management and mitigation strategy.	Architect
Establish quality check schedule and policies.	Architect, Project Manager
Establish quality reporting process.	Architect, Project Manager
Establish and maintain security policies and SOPs.	Architect, Project Manager
Establish IT security reporting processes.	Architect, Project Manager

12 APPENDIX V: DATA QUALITY PLAN

Standard Operating Procedure: Quality Evaluation of Data

1. Purpose

To describe the process for evaluating the data quality of data extractions provided to investigators for research or process improvement.

2. Scope

Data quality evaluation should include the assessment as well as reporting of the completeness, accuracy and consistency of each key metric in a dataset to be delivered to an investigator for research or process improvement (see Table 1).

Table 1. Data Quality Dimensions Determining Fitness for Use of Research Data (Zozus 2014)

Dimension	Conceptual definition	Operational examples
Completeness	Presence of the necessary data	Presence of necessary data elements, percent of missing values for a data element, percent of records with sufficient data to calculate a required variable (e.g., an outcome)
Accuracy	Closeness of agreement between a data value and the true value	Percent of data values found to be in error based on a gold standard, percent of physically implausible values, percent of data values that do not conform to range expectations
Consistency/ Precision	Relevant uniformity in data across clinical investigation sites, facilities, departments, units within a facility, providers, or other assessors	Comparable proportions of relevant diagnoses across sites, comparable proportions of documented order fulfillment (e.g., returned procedure report for ordered diagnostic tests)

Different methods of comparisons can be used to determine the data quality (Table 2). Comparison of data to sources above the top line can be used to identify actual errors in accuracy, while comparison methods below the bottom line can only be used to indicate that discrepancies exist. Comparison methods in the middle can identify discrepancies and help determine if they may or may not be errors. The suggested process is to conduct an aggregate data assessment followed by an individual data assessment. These assessments are further enhanced when using a gold standard or other validated source for comparison when possible.

Table 2: Data accuracy assessment comparison hierarchy (Zozus 2014).

Comparison to a golden standard/ validated source	↑	Accuracy
Comparison to an independent measurement		
Comparison to independently managed data	↓	Partial accuracy
Comparison to an upstream data source		
Comparison to a known standard	↑	Discrepancy detection
Comparison to valid values		
Comparison to validated indicators	↓	Gestalt
Comparison to aggregate statistics		

When evaluating a dataset the values should be assessed for both accuracy and precision by qualitative or quantitative measurement. Accuracy describes how close a value is to its true/actual value and is representative of systematic errors. Precision describes the consistency of the spread of values when repeated and is representative of random or reproducibility errors (Figure 1).

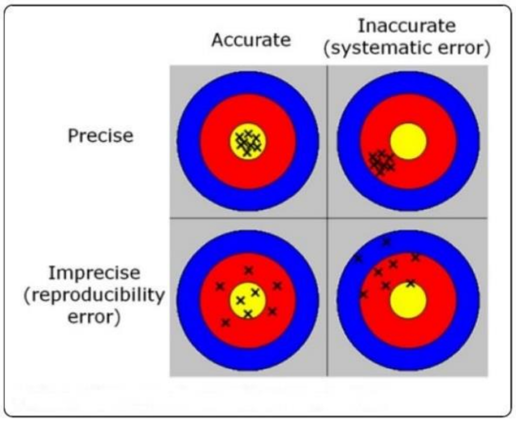

Accurate
Inaccurate (systematic error)
Precise
Imprecise (reproducibility error)

Figure 7: Illustration of accuracy versus precision/consistency (Kellman, Arai et al. 2013).

3. Prerequisites

Create a copy of the dataset and save with 'QA' extension and place in a new QA folder within the investigator's project folder on Emory Box. This is to ensure the original dataset is not altered until a final determination is made to amend data based on the evaluation. The QA folder and evaluation can be saved to the investigator's folder in HIPAA-compliant cloud storage (Emory Box).

4. Responsibilities

Responsible Person/Unit	Action
Quality Analyst (any person(s) who is assigned to conduct the evaluation)	Conducts evaluations for completeness, accuracy and consistency.
Data Analyst	Assists in the conduct of evaluations.
Clinical SME	Assists in the conduct of evaluations.

5. Procedure

Determine completeness, accuracy and consistency of each key variable in the dataset to be delivered to the investigator. Datasets are primarily delivered to the investigator in excel format. Excel format will be the necessary format to conduct the evaluation. The data extraction contained in the excel dataset should be evaluated for quality before delivery to the investigator.

In some cases it may be time-prohibitive to assess each variable. Determine the key metrics to evaluate.

1. Completeness Evaluation
    a. Determine the presence of all necessary variables or columns that are requested by the investigator.
    b. Aggregate value assessment can be done by assessing normal distribution. Create a graph of the count of values over time cross referencing the variable to another appropriate date/time variable in the dataset. If there are discrepancies in the distribution these should be evaluated. The # of reported values should be consistent over time. If there are a large count of missing values it may suggest some data are missing. If there is a larger count of usable values than expected to be present it may suggest duplicate data exists within the dataset. Aggregate value assessment is typically followed by individual value assessment to determine the cause of data incompleteness.
    c. Individual value assessment is assessing if there are null or missing values in the column of each key variable. Create an excel filter to measure presence of necessary data elements, percent of missing values for a data element, percent of records with sufficient data to calculate a required variable (e.g., an outcome). From the drop down menu of the excel sort function review the values to determine if there are '(blanks)' or 'null' values. The nomenclature is not consistent so may be called 'N/A', 'NA', 'blank', etc. Pivot tables are also useful for this assessment instead of reviewing records by sorting.

       There may be instances where the completeness of a variable is distributed across more than a single variable. For example, HIV viral load results may exist as two columns: one column may be a lab result value quantifier (< or >) that exists in a separate column from the lab result value.

      i. Determine if these can be explained.
      ii. Determine if any correction is necessary. Blank values can be considered null values, which may be determined to be usable data. Blank values that are usable may be replaced as 'null' (when appropriate) in order to calculate % completeness.
      iii. Determine if these should remain in the dataset.
      iv. Determine if these should be reported to the investigator as a limitation.
   d. Further individual value assessment can be done when necessary by manual medical record comparison to compare values in the data extraction with values that exist in the Epic record. The central limit theorem for normal distribution extends that a large enough sample size can be representative of the population. Central limit theorem for normal distribution holds for sufficiently large sample sizes, usually n≥30. This allows comparison of source record data with the research dataset without having to review each record (row) in the dataset. The assessment is to determine if values exist in one source but not the other. It is recommended to evaluate n≥30 medical records or as appropriate.
   e. Report the findings.
      i. Report # of missing values and calculate % of missing values.

$$\%\, missing\ values = \frac{\#\, of\ missing\ values\ in\ the\ column}{total\ \#\, of\ values\ in\ the\ column}$$

      ii. Calculate % of sufficient (usable) values. Usable values are defined as not blank, not null and not otherwise coded as null.

$$\%\, usable\ values = \frac{\#\, of\ usable\ values\ in\ the\ column}{total\ \#\, of\ values\ in\ the\ column}$$

2. Accuracy Evaluation
   a. Detection of data errors can be best accomplished through comparison with another validated source of data (gold standard). Possible sources of data other than Epic may be workbench, clarity reports, or LIMS (Beaker or Ultra C). Accuracy has been described as 1) representational adequacy/inadequacy, defined as the extent to which an operationalization is consistent with/differs from the desired concept (validity), including but not limited to imprecision or semantic variability, hampering interpretation of data and 2) information loss and degradation, including but not limited to reliability, change over time, and error(Tcheng 2010).
   b. Aggregate value assessment can be done by assessing normal distribution similar to completeness assessments. In aggregate value assessments the variable can be graphed over time or some other appropriate variable to create a visual representation that can illustrate the presence of outliers. Aggregate value assessment is usually proceeded by individual value assessment to determine the cause of inaccurate data.
   c. Individual value assessment can be done by searching for outliers in the column for the key variable and then reviewing the outlier values in the dataset. In excel from the drop down menu of the sort filter the values can be reviewed to determine if any are

inconsistent. Pivot tables are also useful for this assessment instead of reviewing records by sorting.

i. Determine if these can be explained.
ii. Determine if any correction is necessary.
iii. Determine if these should remain in the dataset.
iv. Determine if these should be reported to the investigator as a limitation.

d. Further individual value assessment can be done when necessary by manual medical record comparison to compare values in the data extraction with values that exist in the Epic record. The assessment is to determine if values are the same in both sources.

e. A clinical SME or biostatistician can aid in determining accuracy. SMEs understand the appropriateness of values in a clinical environment and for data analysis.

f. Report the findings.

i. Report # of inaccurate values and calculate % of inaccurate values.

$$\%\,of\ inaccurate\ values = \frac{\#\,of\ inaccurate\ values\ in\ the\ column}{total\ \#\,of\ values\ in\ the\ column}$$

3. Consistency/Precision Evaluation

a. Assess the consistent representation of data. Representation is the extent to which data are presented in the same or correct format. The goal is that values in a column are uniformly expressed. For example, a common data inconsistency is the expression of date values as mm/dd/yyyy vs. mm/dd/yy vs. mm-dd-yyyy vs mm-dd-yyyy-ss, etc.

b. Aggregate value assessment can be accomplished by creating a graph to help visualize normal distribution that can help to identify the presence of data errors.

c. Individual value assessment can be accomplished by reviewing data with the excel sort function or using a pivot table to evaluate outliers.

d. Further individual value assessment can be done when necessary by manual medical record comparison to compare values in the data extraction with values that exist in the Epic record. The assessment is to determine if values are the same in both sources.

e. A clinical SME or biostatistician can help determine the appropriate expression of data so that it is representative of a true value.

f. Report the findings.

i. Report # of inconsistent values and calculate % of inconsistent values.

$$\%\,of\ \ inconsistent\ values = \frac{\#\,of\ inconsistent\ values\ in\ the\ column}{total\ \#\,of\ values\ in\ the\ column}$$

4. Reporting – Reporting the evaluation is done within the excel table and is also communicated in emails to the HIV Disease Registry data team and the investigator and his/her research team as appropriate. Create a worksheet in the original excel dataset labeled as 'QA' summarizing the results of the data quality evaluation. Based on the report the clinical SME, biostatistician and/or data analyst may determine data to be corrected before delivery to the investigator.

a. Data Limitations are reported in the summary based on findings in the evaluation.

b. Completeness Evaluation

i. Report the # and % of missing values for each key data variable.

ii. Report the # and % of sufficient (usable) data for each data variable.

c. Accuracy Evaluation

i. Report the # and % of inaccurate values for each key data variable.

d. Consistency/ Precision Evaluation

i. Report the # and % of inconsistent values for each key data variable.

6. References

Zozus, M. N. H., W.E; Green, B.B; Kahn, M.G.; Richesson, R.L.; Rusinocovitch, S.A.; Simon, G.E.; Smerek, M.M. (2014). Assessing Data Quality for Healthcare Systems Data Used in Clinical Research (Version 1.0), NIH Health Care Systems Research Collaboratory.

Kellman, P., et al. (2013). "T1 and extracellular volume mapping in the heart: estimation of error maps and the influence of noise on precision." J Cardiovasc Magn Reson 15: 56.

Tcheng, J. N., M.; Fendt, K. (2010). Data quality issues and the electronic health record. Drug Information Association Global Forum.

13 APPENDIX VI: DATA SECURITY PLAN

The primary purpose of this database is for research, population statistics, data mining, and program evaluation for improving health outcomes. According to the HIPAA Privacy Rule, PII used for research must first obtain the patients' consents through a formalized procedure. However, this database is eligible for Waiver of Authorization in which three conditions are met[:

- Use or disclosure of PII involves no more than minimal risk to the confidentiality of the patients. This is accomplished by establishing a compliant security plan to protect individual identifiers from improper use, and documentation that assures PII will not be disclosed to any person or entity outside the scope of the project unless required by law or for regulatory oversight.
- Research cannot practicably be conducted without this waver because it is more than likely that every patient cannot be consented even if attempted.
- Research cannot practicably be conducted without access to this data. Because the database will be used for population surveillance and program evaluation, access to a large sample population of patients is necessary to produce reasonably accurate results.

One of the five steps for the operations security process is the identification of the data that needs to be protected. There are four data sources that are identified for the proposed HIV disease registry that include: Epic EHR, clinical laboratory database, the clarity reports that are derived from Epic, and the existing paper records from which additional information may be extracted and added to the database. All of these sources contain vast amounts of PII. The healthcare system logs over 600,000 encounters per year and the system was implemented in 2009, and additionally includes at least five years of legacy data that was transferred from the old medical records. So the system contains eight or more years of patient health information. While the disease registry may be focused on HIV, it actually accesses every encounter for all patients in the Epic EHR. The linkage between the HIV disease registry and the data sources do not distinguish the information that is aggregated through the ETL processes. The database will contain raw, identifiable data with access to all patients.

To identify and assess appropriate security measures, there must be consideration for all the places data can reside – in motion, at rest, in use, and discarded. This takes into account where data exists in the entire enterprise architecture, which includes operations, physical location, network system, operating system, as well as all associated applications. It is important to note that this database resides on the healthcare organization's informatics enterprise framework, so is subject to the security controls that are in place there, but follows policies of both institutions. Having this on the healthcare organization's server was a strategic advantage for allowing access to PII. It also alleviates the need for an interconnection security agreement for sharing PII with external systems. However, this creates a potential conflict between institutional policies, so a policy or agreement should be produced that can be used to resolve these issues. This will likely occur with the data use agreement between the organizations. The data owner should have the superseding rules; however, Emory seems to share many of the same policies but also exceeds the security rules that is used by the healthcare organization. Because the Emory policies more closely follow HIPAA Section 164.308 and are more comprehensive, the HIV disease registry chose to predominantly base measures on the Emory policies.

The data will be extracted from the sources and aggregated in several places: the HIV disease registry itself, the point at which ETL processes occur, and in limited data sets that are extracted for actual research. An authorized data analyst runs queries on the registry to extract only the information that is needed. Investigators will use this limited data set for various purposes that include research and population statistics. It is expected that this limited data set will be de-identified and then re-identified using a coding system for which a key is created that could link the data. Access to this data is based on roles, and access outside these roles must follow a procedure for approval.

- For data in motion – All data will be housed and will remain on the healthcare organization's informatics enterprise system to take advantage of the security already in place, and to assuage concerns that data stewards may have for data leaving its umbrella of protection and data governance. This is a key point in negotiating access to the data and its use on the proposed HIV disease registry. This will keep data within the controlled environment to provide security to data in motion as it transfers between systems and processes across the network. Data are available for remote access using VPN that protects the connection to the system, but requires an application process with multiple levels of approvals to gain this access. Data encryption is an effective security measure for protecting the data itself.
- For data at rest – Data are stored in several locations, including the HIV disease registry, network servers, as well as on backup systems. Encryption is the primary defense, but physical controls are also necessary to protect from unauthorized access to hardware.
- For data in use – Data are in use on database applications, server applications, with the front-end interface and the SQL interface. Data are also in use in staging areas where it is transformed and loaded once being extracted from data sources. An effective control in Epic is controlling access to printing and copying functions by limiting them to only those roles that require it. Role-based access with functional controls is a good model for the database as well. Another effective security measure in place is access to medical records is limited to computers that are connected with a VPN or on the physical network.
- For data that is disposed –Discarded data on hardware may contain PII or enough elements to be used in combination to link to an individual. Discarded data exist on computers and

servers, and possibly on external media where it poses the highest threat. Controls and policies should exist to prevent data from being stored on portable media, and encryption should be used for data that are stored.

13.1 PII Confidentiality Impact Level Assessment

For the PII Confidentiality Impact Level Assessment the National Institute of Standards and Technology (NIST) method was adapted from 'Standards for Security Categorization of Federal Information and Information Systems' for determining security categorization that are useful for breaking down the components for PII security for this analysis. These formulas are intended for the three security objectives defined by FISMA for information systems: confidentiality, integrity, availability. These were modified to include the factors for determining PII confidentiality impact levels as outlined by NIST: -identifiability, PII quantity, data field sensitivity, context of use, obligation to protect, access to PII, and location of PII.

Security categorization is used to establish a guide for the security of an information system. Security categorization provides assessment of how and where PII exists on the system and suggests the importance of securing this information at the point it exists. Using the NIST standards for security categorization, five individual components were considered for which security could be assessed because of possible exposure to PII. These included the database, application, server, network layers, and backup system that are combined to represent the overall enterprise system for this database.

Based on the NIST formula, the potential impact is considered for each component of the information system, where the composite information system represents the highest value across all layers for each individual factor. The composite security categorization has high impact across the board because there is HIPAA protected patient medical information derived from a direct connection to an electronic health record management system. Because this system resides on the same enterprise as the EHR, the data migration of data remains within the network security protocols providing the possibility to import directly into the database without de-identification. It is this reason that the PII confidentiality impact level is high for every category. It is important to note that the context of use when it is de-identified for research purposes will have low impact as a single limited data set; however, the existence of PII at rest infers high impact for its context of use (aggregation of data containing PII for research, surveillance, and evaluation) while within the database in its raw form. Therefore, because impact is measured by the greatest impact level across all the components it must be considered as high impact. Everywhere else the data exists in raw form. Security management of the database occurs after this data transfer as well as during the data migration with both incoming and outgoing data transfers.

Security Categorization = $SC_{database} + SC_{db\ applications\ and\ OS} + SC_{server} + SC_{network} + SC_{backup} =$

<u>**Composite**</u>

$SC_{information\ system}$ = [(identifiability, high), (PII quantity, high), (data field sensitivity, high), (context of use, high), (obligation to protect, high), (access to PII, high), (location of PII, high)]

Identifiability: Principles of Identifiability of Health Information includes four factors for identifiability[4]:

- Replicability – Data associated with the patients that have high replicability are included in the database. This would include information such as SSN, medical record number, and demographics. These data elements are consistent for each individual, thus providing a link that could be used to identify the individual.

- Data source availability – Data such as address and phone number are publicly available and can be used to identify an individual whose information is stored in the database. If address or phone information were stolen/lost/disclosed by an unauthorized user, then there would be a potential data breach if that data contained information that could be linked to individuals by using public data sources. Therefore, this contributes to a high PII confidentiality impact level.
- Distinguishability – The database and limited data sets will contain enough information that could be used to distinguish individuals when the data elements are used in conjunction. PHI may include medical record numbers, laboratory and admissions dates, and diagnoses.
- Assess risk – Because this data contains replicable, distinguishable data of which some are linkable to available public databases, the risk impact is high.

Quantity of PII: Because this system collects 600,000 encounters a year, the risk impact is high with the potential of affecting thousands of individuals with high associated costs for data breaches or unintentional disclosures.

Data Field Sensitivity: The database contains raw data, which includes PHI in its original form. These data elements infer a high-risk impact because of the nature of PHI in this form.

Context of Use: The final product generated by the HIV disease registry is a limited data set that has PII removed. The limited data set will have de-identified information that is re-identified using a random number for each patient. However, the existence of multiple data sets all for research context using the same random numbers for patients can lead to identifiability. The identification of patients would likely cause harm to the individuals, and disclosure will have legal implications for both institutions. For the purpose of research, the context of use can be considered as high impact.

Obligation to Protect: HIPAA Privacy laws mandate an obligation to protect PII, inferring a high impact level for any disclosure or loss of PII.

Access to PII: PII is accessible by many people that include the informatics team assigned to the database, IT team, and the committees that manage the database. Even though there are access controls and encryption, any disclosure or loss of PII or limited data sets by any of these individuals will result in a high impact on the organization and harm to individuals.

Location of PII: Because there will be a release of numerous limited data sets that will reside in external systems for which the end user is responsible for the control of security, there is an opportunity for identification of patients outside the informatics framework of Emory CFAR and the healthcare organization. There is a high impact level for harm caused to patients by identification and for the damages and cost to both institutions.

13.2 OPERATIONAL SECURITY ANALYSIS

To accomplish a risk assessment, each risk has a matched threat and vulnerability; therefore ongoing analyses for these should be included as part of the operational analysis. Operational safeguards provide assurances for confidentiality, integrity, and accountability for PII. Administrative controls can be used as a security measure to establish the rules governing an informatics environment. These privacy rules include policy, awareness, education, and training for control of behaviors when managing and handling PII. Emory has a program for administrative controls that are divided into nine categories, all of which will be applicable to this database, except when superseded by the healthcare organization's policy for their role as data owners. The following are the Emory measures as required by HIPAA security policies (45 CFR 164.308) for achieving compliance for the database:

1. Security management process [45 CFR 164.308(a)(1)]

2. Assigned security responsibility [45 CFR 164.308(a)(2)]
3. Workforce security [45 CFR 164.308(a)(3)]
4. Information access management [45 CFR 164.308(a)(4)]
5. Security awareness and training [45 CFR 164.308(a)(5)]
6. Security incident procedures [45 CFR 164.308(a)(6)]
7. Contingency plan [45 CFR 164.308(a)(7)]
8. Evaluation [45 CFR 164.308(a)(8)]
9. Business associate contract and other arrangements [45 CFR 164.502(e), 164.504(e), 164.532(d)(e)]

The healthcare organization's policies are not as organized or comprehensive as Emory. The healthcare organization includes a specific policy for HIPAA Information Security Rule (45 CFR 164.316(a)) requiring additional administrative policies and documentation to ensure confidentiality of PII but has a group of general security policies that serves for compliance to HIPAA 45 CFR 165.308. The healthcare organization's Data Security-General Policy includes most of these components, while the others exist within a heterogeneous set of policies.

- Must establish and maintain organizational policies and procedures for compliance with HIPAA Information Security Rule.
- Must establish and maintain organizational policies and procedures to ensure availability, confidentiality, and integrity of PII. Information must be made available when requests are made by patients. In addition, the integrity of data is critical in decision support at point of care to determine diagnosis and/or treatment and care.
- Staff must be informed and trained on all policies and procedures applicable to their role.
- Policies and procedures for system security will be developed in respect to:

    - Size, complexity, and capabilities of Emory.
    - Organization's technical infrastructure, hardware, and software capabilities that incorporate the infrastructure on which the database is housed.
    - Cost of implementing security measures.
    - Based on risk assessment that considers impact level and probability.
- Ensure that policies and procedures are aligned with organizational culture and objectives.
- Conduct an annual review of policies and procedures.
- Maintain written documentation of policies and procedures and retain these for 6 years from the date of creation.

Education, awareness and training provide guidance on the proper handling of PII, sanctions for disclosure or loss of PII, and realization of the risks associated with PII (threats/vulnerabilities). At Emory, the most extensive training for research with human subjects and PII occurs with the Collaborative IRB Training Initiative Program (CITI) that is required every two years. CITI includes several biomedical research and good clinical practice modules related to HIPAA compliance. In addition to this training, Emory also requires completion of training courses in HIPAA Security Awareness and Clinical Research, all of which includes coverage on how to handle and manage PII in addition to imparting awareness for the laws and policies of HIPAA. The HIPAA training also includes individual security responsibilities, knowledge on common security threats and vulnerabilities, best practices, protection guidelines, password management, system security procedures, and how to report security incidents. This is outlined in Emory's Security Awareness and Training Policy. This policy also requires IT staff be aware and trained to comply with login monitoring, audit control and review, data backup, disaster recovery, access management, activity review, password structure, and regulations reminders. Emory administrative safeguards for security awareness and training also include individual policies for security reminders, protection from malicious software, log in monitoring, and password management. The healthcare organization includes HIPAA training for its own staff, but is optional for Emory staff.

13.3 Minimizing the Confidentiality of PII

There are several mitigation techniques for privacy-specific safeguards of PII. The goal is to balance disclosure risk against data security to effectively protect the confidentiality of PII.

- Minimizing the use, collection, and retention of PII

  This is a challenging principle for the purpose of this database since there is yet-to-be-known research parameters and population statistics. Because of its wide scope, the data needs for this system are expansive and contain all types of PII. Minimization of use is controlled by the executive committee whose role is to approve all the research done with data and derivative data collected from this database. This committee will ensure responsible, ethical use of the data. Another mitigation technique is to provide only limited data sets to requestors. For these, all PII are removed as recommended by the safe harbor method. Data use agreements are also required that will limit how the data can be used, and especially limit the use of data to only the agreed upon purpose(s). Sanctions are outlined in the agreement for any violation, and may include legal action and criminal charges.

- De-identifying information

The HIPAA Privacy Rule provides guidance for de-identification and is managed through OCR and at the local level by Emory IRB and healthcare organization's Research Office for Compliance. The Rule permits a covered entity to create a limited data set that is not individually identifiable, or there is no reasonable basis that it can be used to identify individuals, or has minimal risk of this under the Waiver of Authorization. The disadvantage of de-identification is that it could limit the utility of the data. The limited data sets produced by this database will be re-identified, thus creating a random code, which will be linked to the de-identified information. Only the covered entity has access to the key that is the mechanism to relate the de-identified information in the limited data sets with the identifiable information contained in the data sources. This key is never shared with the requestor. Only the covered entity, regulators, and legal action are permitted to use this key.

- Anonymizing the information

Data extracted from the database can be anonymized depending on the purpose or research. During the development of the database, a limited data set was anonymized by removing PII and suppressing data elements to be used to populate an off-site prototype database for testing. The suppressed data omitted portions of records that may have included lab values of certain patients, or clinical data of others. Anonymization of data would also occur for patient of sample population statistics where data are replaced with average values, such as average viral load or CD4 data of a group of patients, and even averages for individuals. This method of anonymizing data would be used to provide data research and statistical analysis.

13.4 Types of Controls

These policy driven safeguards are well outlined in Emory policies and can be applied to the management of this database.

- Technical controls for identification and authentication
    - All users will be assigned a unique ID that will allow identification, tracking, and monitoring of that users activity. This is only effective if the unique ID is not shared with others, so policy must establish rules and sanctions for violation of anti-sharing of ID policy.

  - Employ mutual authentication processes so the data sources will authenticate the database and the database will in turn authenticate the data sources. Multifactor authentication provides protection from impersonation attacks where an unauthorized user may attempt to authenticate him/herself as the database (or data source) so to gain access to PII on the other system.

- Administrative controls for access and authorization
  - Implement role-based authorization where a user is assigned access to data based on their role. For example, this could be used to limit data analysts to querying and viewing data, and preventing them from making any modifications to the data or database applications. This is managed by creating access control lists, which is a simple table mapping individuals to roles and mapping those roles to allow or deny privileges.
  - Ensure there is a policy to immediately disable access to terminated employees. Immediate removal of access will prevent any retaliatory attacks by the terminated employee, or remove the availability of that user ID so that it cannot be hacked or used by an unauthorized individual.

- Technical controls for auditing
  - Auditing is an important tool for identifying and tracking to a source so that all events can be monitored. This allows for monitoring of events, such as attempts to access PII, addition/deletion of users, modifications of permissions, and changes to configuration. According to HIPAA Security Rule 164.312(b), these elements should be logged and audited:
    - User unique ID
    - Login date/time

- Activity time/period
- Description of the event
- Success or failure of the event
- Source of the activity

- Physical control for accountability
    - All storage devices and removable media containing PII should be identified and their locations tracked each time they are moved. The contents should be included in this procedure so in the event of disclosure or loss, the contents of the device or media can be immediately identified for risk/damage assessment. This is particularly important for commonly discarded or re-purposed hardware for backup systems and computers that use storage drives/media.
- Administrative control for accountability
    - Modified data is logged and tracked to the user to corroborate that any changes made are done so by an authorized individual implementing proper procedure. This creates a mechanism to ensure any changes in data stored in the database are authentic and not caused by an intentional or unintentional error when data is in use, in motion, or at rest. For example, this monitoring of authentication will identify if an unauthorized user attempts to alter, add, or delete data. This provides a system for accountability of authentication.
- Administrative – These would be applicable Emory guidelines for the system users and system administrators who are Emory staff. The healthcare organization staff have their own administrative controls.
    - Security management process to include risk analysis, risk management, sanction policy, and system activity review.

- Assigned security responsibility designates an officer to ensure HIPAA compliance and accountability. The principal investigator of the database will likely assume this role as part of the responsibility for management and oversight.
- Workforce security assigns accountability to each role and associated responsibilities to ensure compliance with laws and policies. This will also include a procedure to ensure proper clearance before access is given to any PII, as well as procedure to remove access and responsibility from individuals upon termination.
- Information access management policies establish how role-based access is administered to each staff member, and the rules for how access can be established or modified.
- Security awareness and training will include security reminders to all staff to notify them of any changes to these policies or laws, as well as to alert staff to any potential threats or vulnerabilities. A policy also requires monitoring of access to the information system at any point to ensure security and investigation of questionable activity. Staff will also be trained on password management and policy to ensure adequate security to the information system.
- Security incident procedures establishes how to report and respond to any event that should be recognized in keeping the system secure.
- Contingency plans include data backup procedures, disaster recovery plan, security procedures for operating during emergencies, processes for testing, as well as a criticality analysis to assess the impact to the system during emergencies.
- Business associate contract and other arrangements

  - A critical control is to ensure one-way migration of data from the source to the database. This maintains the integrity of the data in the source, so that any changes made at the database level would not affect the data within the EHR.
  - Another important control is to protect modification of data in the database. For example, if data was manually modified in the database, we do not want a data transfer from the EHR to restore the data back to its original form. This is an important feature to correct apparent errors made in the EHR that can be corrected in the research database.
- Physical – Because the system is housed at the healthcare organization's facilities, theirs would be the appropriate primary guidelines. This is outlined in the healthcare organization Information Access Management Policy. Like administrative controls, the type and extend of access that will be authorized will be role-based and assessed with risk analysis.
  - Facility access control establishes policies to secure the physical location of system hardware.
  - Workstations will be approved for use with PII and will have comply with all appropriate administrative and technical controls such as accessibility and minimum information necessary.
  - Workstation security will require login using credentials that are provided to the database administrative staff for access to PII, determine level of authorization based on role, and use a password authentication.
  - Device and media controls will include maintaining exact copies of data when necessary, adequately destroying unnecessary or no longer needed data using effective sanitation tools, properly destroying removable media and storage devices, ensuring

safe transport of storage media, ensuring data encryption, and maintaining records of these devices and media and logging location and relocation.

- Technical – Similarly to the physical controls, the healthcare organization's policies should be the primary guidelines since this system is housed on their enterprise system.
    - Access controls include assigning unique user IDs, emergency access procedures, automatic logoff when a computer is left for a specific amount of time, and encryption/decryption. This also includes policies surrounding remote access, which is available for Epic EHR.
    - Audit controls to review systems for unauthorized disclosure or security breach.
    - Integrity procedures ensure authentication measures are engaged, and the appropriate level of security is used. These measures ensure data is stored and transferred in its intended form without modification.
    - Person/entity authentication is required for Emory to authenticate all database users before access is granted.
    - Transmission controls minimize the risk of unauthorized access or modification of PII during transfer between data sources and the database. This would include the transfer of PII or limited data sets in emails or between systems.