A Guide to Using Data from EPIC, MyChart, and Cogito for Behavioral, Social and Systems Science Research

Authors:

- Eric Ford, *PhD* [1]
- Julia Kim, *MD MPH* [2]
- Hadi Kharrazi, *MD PhD* [1, 2]
- Kelly Gleason, *BS* [2]
- Diana Gumas, *MS* [2]
- Lisa DeCamp, *MD MSPH* [2]

[1] Johns Hopkins School of Public Health
[2] Johns Hopkins School of Medicine

Prepared for:

Johns Hopkins School of Medicine
Institute for Clinical and Translational Research (ICTR)
Behavioral, Social and Systems Science (BSS) Translational Research Community (TRC) Advisory Board

Apr 2018

TABLE OF CONTENT

■ EXECUTIVE SUMMARY

In Jan 2017 the Institute for Clinical and Translational Research's (ICTR) Behavioral, Social and Systems Science (BSSS) Translational Research Community (TRC) advisory board funded this project to examine the availability of social and behavioral data in JHMI's EPIC EMR/PHR systems. Both researchers and administrators recognize that patients' social determinants play a critical role their care experiences and outcomes. Being in Maryland with it global budgets and population-based reimbursement scheme, it is advantageous for JHU/JHMI to find cost-effective, community-level solutions that improve the population's health status. The vision of the BSSS TRC board is to enable JHU researchers to utilize social/behavioral data collected from JHMI patients and stored in various data sources such as EPIC.

In the first phase of this project (3 months), the project team developed a guide that can be used by JHU researchers to understand: 1) different types and frameworks of social and behavioral data; 2) learn from current and previous attempts to extract social/behavioral data from EPIC at JHMI: and 3) explore some aspects of the common social and behavioral data captured in EPIC. EPIC data elements that are available and JHMI processes and procedures are evolving. This guide is meant to capture current state and parts of it will be transitioned to a webpage in the second phase of the project to allow for timely updates.

The first phase of this project also produced a detailed proposal for the second phase of the project that will involve a more in-depth analysis of social/behavioral data captured in JHMI's EPIC (12 months).

This guide provides the following sections:

- *Quick Guide*: provides a high-level picture of how different data requests to extract social/behavioral data from EPIC are managed.
- *Background*: reviews the current literature and frameworks proposed by various institutes and researchers to assess various health determinants including social/behavioral data.
- *Methods and Results*: offers three approaches and results to understand the complexity of social/behavioral data extraction from EPIC:
    - An environmental scan to explore the efforts of various EMR vendors in collecting and organizing social/behavioral data.
    - Expert interviews to reflect on the experience of JHU researchers and staff who have extracted social/behavioral data from EPIC EMR/PHR.

    - A proposed data and human matrix that will be used to code/tag existing social/behavioral data in EPIC in the next phase of the project
- *Appendices*: additional details about the interviews, extracting data from EPIC, and sample data matrixes applied to common social/behavioral data

The next phase of this project will start in Jul 2018. Please contact Dr. Hadi Kharrazi (kharrazi@jhu.edu) for activities planned for the second phase of the project. For all other information please contact Kelly Gleason (kgleaso2@jhmi.edu).

■ QUICK GUIDE ON DATA RETRIEVAL

Collecting and extracting social/behavioral data in EPIC for research purposes is a complex task and can be executed in various ways. Often small data collection or extraction efforts are performed manually while larger data collection or extraction efforts require coordination with JHU's Data Trust and the Center for Clinical Data Analysis (CCDA). Regardless of the size of the data collection/extraction, all research-driven data retrievals should be reviewed and approved by local IRBs before any attempts are made to extract data from EPIC.

• *Primary Data Collection*

Please contact CCDA and EPIC's MyChart (PHR) team if you wish to deploy a new questionnaire or generate a new field in EPIC/MyChart for the collection of new social/behavioral data.

• *Secondary Use of Data (Data Extraction)*

- *Data Collection or Extraction*

A manual process might be the preferred method for collecting and/or extracting data for a small sample size (fewer than 100 patients) in EPIC; however, larger numbers are limited due to existing HIPAA liabilities and simply not being pragmatic. Please contact the JHU Data Trust and CCDA if you require larger data cuts that need automated mechanisms (e.g., retrospective queries, real-time retrieval).

- *Data Queries / Extraction Modes*

Depending on your research timeline, you may need different granularity of data:

- **Hypothesis Generation (Exploration)**: You can use this guide to get an overall picture of social/behavioral data captured in EPIC (to be completed at the end of phase 2), or use the Slicer Dicer tool, embedded in EPIC, in order to explore structured variables captured in EPIC.
- **Feasibility Assessment and Proposal Preparation (Estimation/Counting)**: You can use EPIC's Slicer Dicer tool to define your population denominator of interest and perform basic counts prior to obtaining IRB approval. Please contact CCDA if you need help with executing advanced counts.
- **Extracting Data and Building Analytical Files (Extraction/Querying)**: Depending on the size of your population, you may need to approach JHU's Data Trust and/or CCDA to extract the data required for your research. CCDA offers free consultation to provide you with an estimated cost associated with such a data retrieval effort. CCDA can also provide you with data quality checks.

- *Data Analysis*

Epic analytic extracts are often complex and require drawing on numerous fields in various parts of the system, significant data cleaning and preparation in advance of analysis. Researchers will need to work with their own team of statisticians and data analysts / managers to ensure that the appropriate fields are being queried bearing in mind that comparable clinical measures may have different field labels depending on where the information was gathered. Currently, CCDA does not provide such services without reimbursement although there are consultation services that are supported. Both CCDA and ICTR can provide you with contacts for research teams that have previously worked with EPIC data to guide estimates for data / statistical team effort and experience as well as feasible project timeline.

■ INTRODUCTION

The evolving delivery models and alternative payment programs (APMs) that provide incentives for delivering high-quality care are changing both the idea and measurement of value in health care. The goal of the ongoing shift away from fee-for-service systems is to promote improved care for populations at a reduced cost. The common thread to both new models (e.g., Triple Aim) and incentive schemes (e.g., Accountable Care Organizations and Patient Centered Medical Homes) is a focus on prevention.

The primary challenges to measuring prevention efforts are that many of the relevant factors exist outside the health system and risk exposures may occur years before a classically classified disease manifests itself. Moreover, to interrupt the etiology of many diseases, it requires interventions at the behavior, socio-economic and environmental nexus. Taken together, these social determinants play a significant role in the disease types and acuity-levels patients present at the time of clinical encounter. However, relatively few measures related to the social determinants of health are routinely collected in a structured, analyzable fashion, especially in the healthcare EMRs. Therefore, building a framework that identifies the social determinant measures that would be useful in health system administration and clinical research is a necessary first step to realizing the proposed reimbursement models' aims. A natural follow-on activity is to assess the health system's current capabilities and potential capacity to collect health determinant measures.

There is a wide range of personal, social, economic, and environmental factors that influence a patient's health status and care outcomes known as "determinants of health". Increased efforts to provide holistic care and prevent episodic events by intervening on social determinant factors are gaining traction. In particular, value-based purchasing and population health management initiatives are creating capitated payment systems that promote early interventions at the health determinant level. Thus, collecting and analyzing determinant of health measures is increasingly critical for both research and operational reasons.

Electronic medical records (EMRs) used by clinicians may capture some health determinant measures in a structured format. However, many social determinant variables that influence care outcomes are not routinely captured in the EMR or appear in the clinical notes as unstructured narratives. Other social determinants may be entered directly by the patient in a personal health record (PHR) or gathered through surveys related to particular research endeavors. The net effect of collecting, or failing to collect, social determinant measures through these various channels is that researchers

and managers may have difficulty accessing valuable data in a timely and usable form. Understanding the current state-of-the-art in collecting social determinant measures is an important step in building a learning health system that can address population-level outcomes.

The purpose of this guide is three-fold:

- First, the literature describing the different types and frameworks of social and behavioral data are reviewed and synthesized.
- Second, key informant interviews are conducted to learn from current and previous attempts to extract social/behavioral data from EPIC at JHMI
- Third, a plan is produced for exploring some aspects of the common social and behavioral data captured in EPIC.
- These efforts represent phase one of a two-part project. The second phase will document the social determinant measures available, provide a guide to researchers wishing to integrate them into studies, and make recommendations on how to integrate new measures going forward. Additionally, a transition of the guide to a web resource is proposed to promote timely updates when there are changes to available measures or JHMI processes or procedures for obtaining EPIC data.

■ BACKGROUND

• *Literature Review of Existing Frameworks*

The National Academy of Medicine (NAM) Committee on the Recommended Social and Behavioral Domains and Measures for Electronic Health Records has chosen three frameworks, "The Public Health Model of the Social Determinants of Health", "Pathways Linking Socioeconomic Status and Health", and "Multilevel Approach to Epidemiology" to guide their report "Capturing Social and Behavioral Domains in Electronic Health Records." [1]. Following is a review of the literature describing these three frameworks:

○ *Public Health Model of the Social Determinants of Health*

The "Public Health Model of the Social Determinants of Health" (Figure 1) describes the relationship among social determinants, health care system attributes, health outcomes, and disease-inducing behaviors [2]. The model demonstrates the relationship between social determinants and health through its structure, and the nature of causal relationships between social determinants and health through analyses that it facilitates.

The model states three components of social determinants that are well-established in the literature: (a) socioeconomic conditions, (b) psychological risk factors, and (c) community and societal characteristics. Socioeconomic determinants include age, sex, and education. There is an empirically demonstrated causal relationship between socioeconomic status and health [3]. Psychosocial risk factors include social support, self-esteem, chronic stress, and isolation. Psychosocial factors are increasingly recognized for their influence on health [4]. Community and societal characteristics include income inequality, social capital including civic involvement, and level of trust. Studies have demonstrated the link between health outcomes and social support, social networks, and social isolation [5, 6].

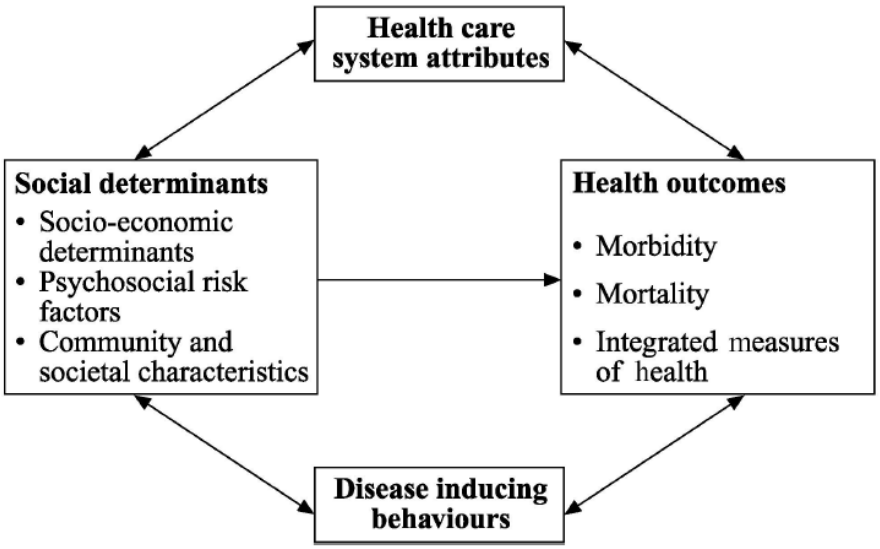


Figure 1. Public health model of the social determinants of health

The public health model depicts that certain health care system attributes are connected to population health and care inequalities. Primary care is associated with improved health outcomes [7]. Socioeconomically disadvantaged groups are less likely to use preventive measures, including immunizations, dental services, and antenatal care [8]; however, it is still unknown whether the lower use of preventive services in these groups is the result of less access, less information, or more pressing priorities. Disease inducing behaviors explain only a small proportion of the effect of social factors on health outcomes [9]. This model suggests that psychosocial processes influence the ability to initiate and maintain health-enhancing behaviors [10]. Constraints including extended exposures and persistent poverty may lessen the effectiveness of health promotion efforts in disadvantaged individuals and communities.

The public health model provides the opportunity to include individual level variables and ecological level measures in the same analysis. It demonstrates a framework for understanding causal pathways between social determinants, disease inducing behavior, health care systems, and health outcomes.

- *Pathways Linking Socioeconomic Status and Health*

"Pathways Linking Socioeconomic Status and Health" is a simplified model developed by the MacArthur Research Network on Socioeconomic Status and Health to depict pathways linking socioeconomic status and health (Figure 2). The MacArthur Research Network on Socioeconomic Status and Health aimed to identify mechanisms by which disadvantaged individuals develop poorer health due to socioeconomic status.

The pathways model posits that health care is an important pathway but acknowledges that access to health care alone will not eliminate health disparities, but rather, may work in tandem with improved social conditions to provide disadvantaged groups with better health outcomes [11]. Environmental exposures are included in the pathway linking socioeconomic status and health; low socioeconomic status communities are both subjected to more environmental hazards and have access to fewer resources to mitigate these hazards. The model acknowledges that socioeconomic status patterns leads to certain health behaviors, which contribute to higher morbidity and mortality. Authors of the model list smoking as a key health behavior that differs by socioeconomic status and contributes to health disparities. Allostatic load is included in the model as a measure that captures the biological consequences of stress. There is evidence that the chronic stress, sometimes referred to as toxic stress, is associated with lower socioeconomic status results in higher allostatic load, which is associated with increased vulnerability to disease.

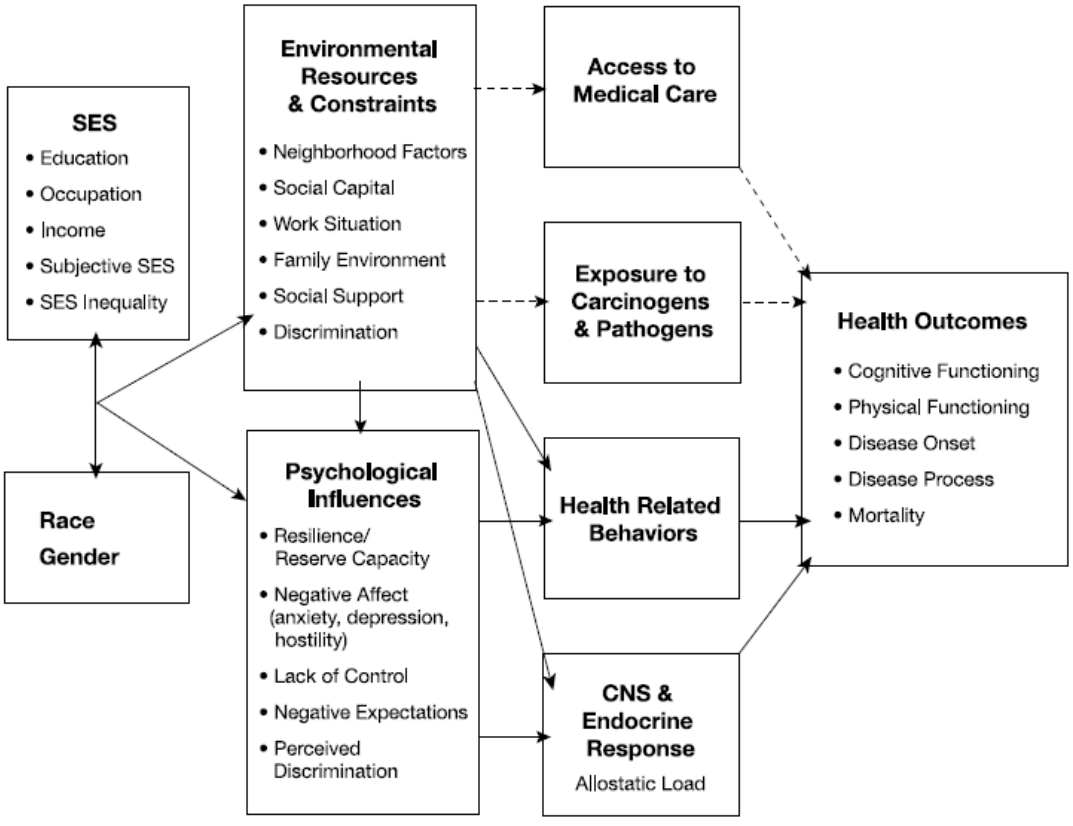


Figure 2 - Pathways linking socioeconomic status and health

- *Multilevel Approach to Epidemiology*

The "Multilevel Approach to Epidemiology" displays that there is no single explanation for the heterogeneity across health outcomes (Figure 3). The model depicts that while individual risk factors, genetic factors, and pathophysiological pathways are important to realizing differences in health outcomes across groups, they must be viewed through a larger lens that takes social and economic policies, institutions, and neighborhoods and communities into consideration. Focusing on molecular etiologic forces located within the individual, as genomics does, will not explain the disparities in health by social groups and places. The public health model is largely rooted in this model, though the public health model goes further in demonstrating that social determinants affect health in multiple ways across the life course, both directly through behaviors, and through interactions with the health system people use.

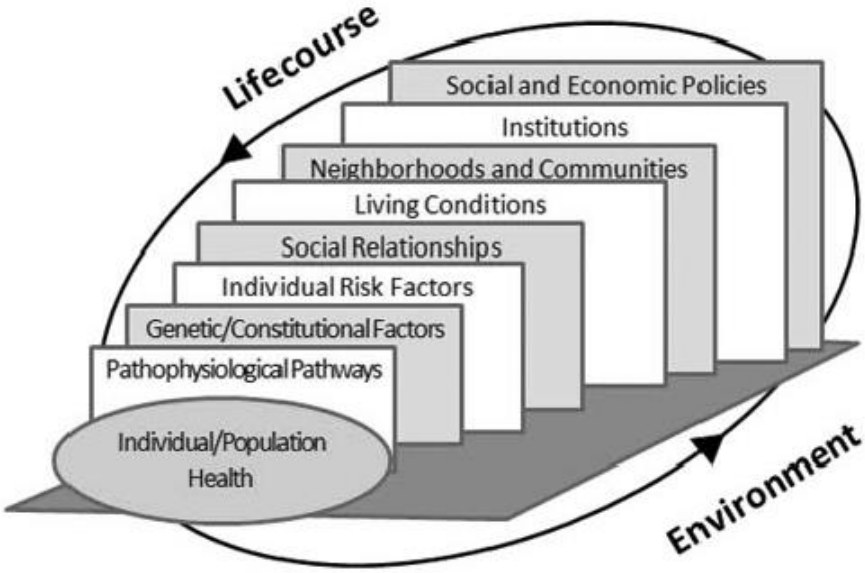


Figure 3 - Multilevel approach to epidemiology

- ***Social Determinants of Health***

The Office of the National Coordinator's (ONC) Meaningful Use 3 program defines five domains to group factors that make up social determinants of health in the US as: sociodemographic, psychological, behavioral, individual-level social relationships, and neighborhoods and communities. Literature examining these five domains was reviewed and synthesized.

- *Sociodemographic*

Throughout history, impoverished individuals have been disproportionately affected with disease burden and have had shorter life spans [12]. Low socioeconomic status is associated with health outcomes [13, 14]; and, individuals of lower socioeconomic status have an increased prevalence of functional difficulties and poor health [13-16]. Financial strain is also a predictor of nursing home placement [17]. Food insecurity is associated with poor physical health status [18] and not receiving home health visits or having a primary care provider [19]. Older adults who experienced food insecurity have reported limitations in activities of daily living 14 years earlier than older adults who did not experience food insecurity [20]. Food insecurity is also correlated with cost-related medication underuse and comorbidities including diabetes and heart disease [21-24]. Education is an established major indicator of socioeconomic status and a risk factor for poor health outcomes [25]. Lower levels of education are associated with an increased risk of cardiovascular diseases and events [26].

Functional difficulties disproportionately impact marginalized populations. A recent examination of active life expectancy found that older black women are disadvantaged compared to their white counterparts in proportion of years expected to be lived without disability [27]. Sex-based differences in clinical outcomes from treatment are well documented [28-30].

- *Psychological*

The link between psychological factors and health is increasingly recognized. Depression and anxiety are both commonly reported and interrelated. Even mild, subclinical levels of depression and anxiety can increase the risk of other diseases including cardiovascular disease, diabetes, and stroke [31-34]. Findings suggest major depression is the second leading cause of disability worldwide [35]. Social status impacts stress, the body's response to demands and threats [36]. Chronic levels of stress have been linked to poor health outcomes, including hypertension and a greater susceptibility to infection [37]. For example, there is evidence that household stresses, including noise, fear of eviction, residential instability, and lack of control, are associated with increased asthma attacks [38].

- *Behavioral*

Disease-inducing behaviors including smoking and alcohol use lead to poor health outcomes [9], while an individual's willingness to change behaviors is linked to improved health. Smoking and alcohol use have a causal relationship with poor health outcomes, including increased mortality [39, 40]. Patient activation is a significant

predictor of health care utilization, patient outcomes, and health behaviors [41, 42]. Patient activation has been associated with positive health outcomes among adults with chronic illnesses [43, 44]. Findings from recent studies suggest that patients with a higher activation are more likely to adhere to medical regimens and effectively manage chronic medical conditions, and less likely to be hospitalized [45-47]. Limited prior studies of patient activation in older adults (individuals age 65 and older) indicate that higher patient activation scores are associated with higher functional status, health care quality, and adherence in older adults [48]. Readiness to change may impact individual's level of engagement in health interventions and, consequently, the success of these interventions [Rose & Gitlin, in-press].

- *Individual-level social relationships*

Individual-level social relationships are associated with health outcomes. Living arrangements are identified as a major indicator of social support as living arrangements facilitate social support [49]. Living alone is associated with poor health outcomes including increased mortality and hospitalizations [50-54] and marriage has been linked to decreased mortality [55-57]. Insurance coverage is associated with improved health outcomes [58-60].

- *Neighborhoods and communities*

There is evidence that neighborhoods impact health independent of individuals' own socioeconomic status. For example, individuals living in lower socioeconomic status neighborhoods have poor health independent of their own socioeconomic status [61-63]. Ease and safety of exercising and availability of healthier foods such as fresh fruits vary across neighborhoods that differ by socioeconomic status and constrain healthy behaviors [63].

■ METHODS

Multiple methods were used to gather information about the extraction and/or collection of social and behavioral data from EMRs. First, an environmental scan was performed to identify the efforts that major EMR vendors have taken to collect and organize social and behavioral data (often delivered as population health management functionalities). Second, several interviews were conducted with Johns Hopkins researchers with significant prior experience with using EPIC for BSSS research and extracting social and behavioral data from EMRs (specifically the Johns Hopkins EPIC EMR). Also, Johns Hopkins websites were searched for information on resources available to support BSSS research, based on interviewer recommendations. Finally, a data matrix (a.k.a., meta-information template) and a human matrix were developed that can be used to tag and categorize social and behavioral data available in EPIC. These matrices can be used to provide researchers with a snapshot of underlying information associated with common social and behavioral data extracted from or collected in EPIC.

• *Environmental Scan*

A review of the EMR vendor websites identified in the "AHA Health Information Technology Survey Supplement" was undertaken to identify the electronic published descriptions of the availability of social determinants measures in the EMR. The researchers went to each vendor's homepage and searched in two fashions. First, a review of the broad functionalities highlighted by the vendor was undertaken. Particular attention was paid to functionalities focused on "population health", "accountable care organizations", or "patient centered medical homes". Next, the term "social determinants" was used to search within the vendors' websites.

• *Expert Interviews*

Several JHMI/JHU staff and researchers were identified and interviewed to understand their experiences with using EPIC social determinants data and their recommendations for designing and conducting research studies using EPIC data. Initial interview participants were identified based on identified content area expertise from the perspective of guide authors. Additional interviews were conducted based on recommendations from participants regarding other researchers who also could provide key information for use in the guide. A semi-open questionnaire was used in interviews either conducted face-to-face or via phone. The interviewees included: Diana Gumas, the leader of Center for Clinical Data Analytics (CCDA) and EPIC Research; Dr. Peter Zandi, a clinician who uses EPIC for social and behavioral research; Mrs. Valerie

Smothers, an expert in the Data Trust Council; and, Dr. David Thiemann and Bonnie Woods, experts in extracting data from the EMR. Interviews were aimed to learn from the interviewees' experiences using the EMR for research. The full notes of the semi-structured interviews can be found in the appendix.

- ***Analyzing Existing Epic Data***

This section describes key stakeholders and their roles in guiding the use of EPIC data for researchers in behavioral, social, and systems science. Stakeholders comprising the human matrix around EPIC data use for research include various data custodians and specialists, data collectors, researchers, and patients. In addition to eliciting information from the interviews as described above in regards to the roles of the various stakeholders in the process of using EPIC data for behavioral social and systems science research at Johns Hopkins, we also collected information available from the Johns Hopkins website and the CCDA performed some preliminary analysis of common social/behavioral data types by applying the data matrix to each of them. In phase 2, we plan to conduct additional surveys to further define stakeholder roles, explain more details about the data extraction process, and apply the data matrix to more social/behavioral data types.

■ RESULTS

• *Environmental Scan*

Table 1 provides a summary of EMR vendors and population health management functionalities, encompassing various social and behavioral data management functions. Overall, based on their websites, EMR vendors do not appear to be prioritizing building functionalities to collect social determinants. To the extent that such capabilities are being developed, they are focused on reducing readmissions and other "downstream" determinants related to reimbursement programs (e.g., ACOs). The vendors do make references to "predictive analytics", but those algorithms appear to rely on existing EMR data fields. While the vendors are not leading the push for including social determinants in EMR systems, the academic literature is growing. Currently large gaps exist in the functionality of social and behavioral data in EMRs. The emergence of groups such as the EPIC Social Determinants of Health Braintrust, may eventually improve and build functionality around SDH measures.

The number of articles in PubMed with a reference to "Social determinants" has grown from 295 in 2000 to 2,197 in 2016. In addition, professional organizations are beginning to recognize the value of collecting social determinant information through the EMR. In particular, the American Academy of Nurses (AAN) has called for capturing social determinants in the EMR [64]; however, AAN recognized that the terms, variables and fields needed further development.

Table 1 - EMR vendors and population health management functionalities (encompassing various social and behavioral data management functions)

Vendor	Website URL	Functionalities	Social Determinant	Comment
Allscripts	http://www.allscripts.com/market-solutions/hospitals-health-systems	Modules for: (1) Population health management; and, (2) Precision medicine	Refers to white papers on building a successful ACO.	Population health module primarily aimed at care coordination.
CPSI / Healthland	http://www.healthland.com/solutions/hospital/inpatient_ehr/	Standard EMR modules. No mention of Population health, etc.	No results	A relatively sparse set of functionalities
Cerner	https://www.cerner.com/	Module for: Population health management that included care coordination and wellness	No results	Among the systems reviewed, Cerner seems to be paying the most attention to the topic.
EPIC	http://www.epic.com/software	Modules for: (1) Patient engagement; and, (2) Population health management	No results	Population health module primarily aimed at care coordination.

GE / Centricity	http://www3.gehealthcare.com/en/products/categories/healthcare_it/electronic_medical_records/centricity_emr#tab	No mention of Patient engagement or population health management	No results	No mention of any concept related to social determinants.
Meditech	https://ehr.meditech.com/ehr-solutions/hospitals-health-systems		Returns two whitepapers one discussing 'Big Data' and the other focused on 'Population health'. The latter mentions an American Academy of Nurses' call for including social determinants in EMRs.	

- ***Expert Interviews Summary***

Through our expert interviews and review of Johns Hopkins website information, we identified key JHM Resources to support behavioral and social science research, important steps for researchers to consider when obtaining EPIC Data, and challenges to using EPIC data for BSSS research.

- *Behavioral, Social, and Systems Science (BSSS) Community*

The Behavioral, Social, and Systems Science (BSSS) community is designed to create an academic home and collaborative community for diverse scientists from across Johns Hopkins University who are conducting research in the areas of health and behavior, biopsychosocial interactions, social and cultural factors in health, health systems and health services, health IT, and methodologies. The BSSS Community serves as a catalyst to stimulate highly innovative researchers and research programs that expand the translation and dissemination of this research, and facilitate new methodologies for solving current health systems, community, and population level challenges, through systematic interdisciplinary approaches.

Key stakeholders in behavioral, social, and systems science research include: Peter Zandi, researchers in the JHSPH Department of Health Behavior and Society, clinical researchers, and leaders in the BSSS Translational Research Community (TRC).

- *Data Trust Council and Analytic Teams*

The Data Trust Council (DTC) governs JHM data (data in JHM clinical, health plan, and business systems), making such data readily available for appropriate use while protecting patient privacy and maintaining data security. The DTC has subcouncils, each with a different responsibility (*e.g.*, research use, quality improvement, security), to

review and approve data requests and propose policies. The actions and oversight of the DTC were authorized in 2016 when the participating JHM provider entities (including JHH, Suburban Hospital, Sibley Memorial Hospital, Howard County General Hospital, and JHCP) and health plans signed the JHM Data Trust Policy, establishing the DTC and giving it authority to oversee JHM data use and approve data requests.

All Hopkins data, even if not subject to Data Trust oversight (*e.g.*, data collected solely for research, not used for patient care, and not stored in any clinical system), must still be stored, used, and disclosed in compliance with the appropriate agreements regarding data use as well as IRB and Johns Hopkins IT policies and requirements, which include encryption, server security, and access controls.

The "Data Trust Research Data Subcouncil" develops policy and reviews requests for *research* uses of JHM data. Hopkins IT and security experts, working with the "Center for Clinical Data and Analytics" (CCDA), help the Data Trust Research Data Subcouncil assess technical security, access controls, and Deidentification protocols for specific projects. The organizational chart for the Johns Hopkins Data Trust Council can be found in Figure 4 and the Data Trust Analytic Teams within the Data Trust Operations Team can be found in Figure 5.

Figure 4 – Organizational chart of the Johns Hopkins Data Trust Council

Figure 5 – Data Trust teams

The Operations Team is a central team that will support the development of shared Data Trust infrastructure and coordinated analytics. It will play a coordinating role

across the 10 approved Analytic Teams. Analytic Teams work to coordinate analytic efforts across Johns Hopkins Medicine within a defined scope. They help reduce redundant efforts and encourage use of common infrastructure. Analytic Teams also play a role in building data flows to efficiently support analytic needs. These teams will consider and fulfill quality, operational and research-related requests for data. The teams focus on:

- Ambulatory operations
- Ambulatory quality
- Hospital quality
- Hospital operations
- Hospital utilization management
- Finance-integrated analytics
- Population health
- Research/Center for Clinical Data Analysis (CCDA)
- Technology Innovation Center
- Planning and market analysis

Follow these links to access additional information about the Data Trust and see guidelines for requesting access and data.

- Operations and guiding principles
- Data Trust policies
- Requesting access to the Data Trust infrastructure
- Requesting data from an Analytic Team

Analytic Teams approve access to components of the Data Trust Infrastructure for analysts working within their purview. They also consider and fulfill quality, operational and research-related requests for data. Many Analytic Teams operate virtually and may report to different individuals. Below is a list of the Analytic Teams:

http://intranet.insidehopkinsmedicine.org/data_trust/analytic_teams/

Not all research requests for JHM data require review. Many smaller projects (<500 records with acceptable data security plan in the IRB application) are not required to be reviewed by the data trust. (see appendix for Data Trust Review of Research Data Requests FAQ)

- *Institute for Clinical and Translational Research (ICTR)*

The Johns Hopkins Institute for Clinical and Translational Research (ICTR; http://ictr.johnshopkins.edu/about-us), established in 2007, is one of more than 60 medical research institutions working together as a national consortium to improve the way biomedical research is conducted across the country.

The ICTR addresses obstacles in translating basic science discoveries into research in humans, translating clinical discoveries into the community and communicating experience from clinical practice back to researchers. The ICTR houses three Translational Research Communities for investigators across multiple disciplines that focus on drugs, biologics, vaccines and devices; biomarkers and diagnostic tests; and behavioral, social and systems interventions. These communities of researchers help prioritize clinical problems in need of new treatments, apply new technologies and methodologies, support junior investigators, work with translational partners outside of Johns Hopkins, fund pilot projects, provide regulatory assistance and promote efficient research. Another ICTR program, The Research Studio, provides both a place and a process for investigators and their teams to obtain multidisciplinary guidance to solve clinical and translational research problems. Additionally, the ICTR provides research teams across the university and affiliated research institutes with a range of services within five coordinated Cores:

- Translational Laboratories
- Human Subjects Research
- Quantitative Methodologies
- Clinical Research Informatics
- Research Participants and Community Partnerships

Two groups of interest within the ICTR include the ICTR Data Managers Interest Group, and the ICTR Advisory Board and Best Practices Working Group, both run by Kelly Crowley.

- *Center for Clinical Data Analysis (CCDA)*

The Center for Clinical Data Analysis (CCDA) assists researchers with accessing clinical data for research purposes. Services include:

- Preliminary, anonymous data for feasibility, grant applications and statistical sample-size estimates
- IRB-approved case-finding–for study enrollment (mailings, phone solicitation), chart review, and cohort/case-control studies

- Research data extracts – monthly/quarterly integrated extracts from EPIC, EPR, Sunrise/POE, and CaseMix/Data Mart
- More information about CCDA can be found at: http://ictr.johnshopkins.edu/clinical/clinical-resources/clinical-research-informatics-core/center-for-clinical-data-analysis-ccda/

The CCDA is staffed with experienced data analysts who will assist you with access to data while also helping you comply with Data Trust privacy and security regulations. The contact person at CCDA is Bonnie Woods, IT Senior Program Manager, Bonnie.Woods@jhu.edu.

- *Other Data Specialist Resources*

In addition to the data specialists in the ICTR and CCDA, other data specialist resources include:

- Claire Twose, Associate Director of Research Services, Welch Medical Library
- Bonni Wittstadt, GIS specialists
- Jen Darragh's replacement as social data specialist at Sheridan

- *EPIC MyChart Committee*

The Johns Hopkins EPIC MyChart Committee, led by Steve Klapper and Michele Lang, can serve as a valuable resource in providing operational, clinical, and research perspectives on data collection of social and behavioral variables through MyChart.

- *Institutional Review Board (IRB)*

The Johns Hopkins IRB also serves as an integral stakeholder in the research and the use of electronic record based patient data for BSSS research.

- *Patients*

Patients can also hold an important role in the process of accessing and using EHR-based data for behavioral, social, and systems science research. Community Advisory Boards(CABs) or Patient and Family Advisory Councils (PFACs) can help to identify and refine research questions. The Patient and Family Centered Care Community (PFCC) is an established PFAC collaborative at Johns Hopkins Medicine. The PFCC is run by the Armstrong Institute and was designed to provide a structure for PFACs across varied healthcare settings at Hopkins to work together to ensure that patient/family perspectives and priorities are adequately represented to inform research and healthcare improvement.

How to Obtain EPIC Data for BSSS Research

1. **<u>Formulate your specific research question and data request</u>**

The first steps that a researcher should take to obtain data from EPIC, involve thinking carefully about what data are needed.

- *Define your patient population: For what patients do you desire the data? (e.g. all patients for which I am the PCP, or all patients who meet a set of inclusion and exclusion criteria approved by the IRB, or all patients consented to my study and actively on study in the Clinical Research Management System.*
- *Define your time frame: For what time frame to do you desire the data?*
- *Define your location: From what locations do you desire the data? (e.g. Johns Hopkins Hospital? Bayview Medical Center? Johns Hopkins Community Physicians? Sibley Memorial? Suburban Hospital? Howard County General? All of the above?)*
- *Define your data elements: Which data elements do you desire? (e.g. race and ethnicity, year of birth, smoking status, diagnoses, etc.). It helps a great deal to partner with a physician who actively uses EPIC who can help you take screen shots of data elements that are more unusual.*
- *Contact CCDA: Ask CCDA for an estimate of the cost for a programmer to extract these data for you so that you can then seek funding if needed.*

2. **Consider the following examples of well-structured requests for data as templates for your data requests:**
    - **Example 1**: Adult patients (ages >= 18) seen as outpatients at Bayview and JHH psychiatric clinics from October 1, 2016 to April 30, 2017 diagnosed with major depressive disorder, bipolar disorder, or schizophrenia (either as an encounter diagnosis or on the problem list) having a smoking status that is not "Never".
        - This example answers the question "which patient", what encounter type (outpatient vs. inpatient), what encounter location (specific Bayview and JHH psychiatric clinics), what time frame, and other criteria (diagnoses and smoking status).
    - **Example 2**: All patients with an in-person (outpatient) visit to a Johns Hopkins internal medicine, family medicine, pediatric, psychiatric, pediatric psychiatric or obstetrics/gynecology clinic from April 1, 2013 until July 1,

2016 whose clinician completed the depression screening flowsheet during that visit. See Appendix A for complete list of departments to include.

3. **How to Submit a Request to CCDA**
    - Submit a request for CCDA services using our new iLab application: https://johnshopkins.corefacilities.org/service_center/show_external/3796
    - You will receive an email response, usually within one to two business days.
    - You will receive 2 free hours of service underwritten by the ICTR. This usually covers an initial meeting to discuss your request in detail, a feasibility assessment, a written specification, and an estimate of hours to complete. Please note that it will usually take a minimum of 1 week to have an analyst available to start work. We will communicate the start date when we deliver the estimate of work.
    - Requests for guidance or for simple patient counts may be able to be completed within the 2 free hours. More complex counts may extend beyond the 2 free hours.
    - Data extraction projects usually range from 8 hours to 150 hours depending on complexity. The average project is about 30-35 hours.
    - The cost for CCDA services is $84/hr for standard services and $100/hr for senior analyst consulting.
    - Contact Info: Bonnie Woods, IT Senior Program Manager, Bonnie.Woods@jhu.edu
    - More information: http://ictr.johnshopkins.edu/clinical/clinical-resources/clinical-research-informatics-core/center-for-clinical-data-analysis-ccda

4. Consider Using Slicer Dicer Tool to Explore Preliminary Hypotheses

- *Slicer Dicer – Overview*

Slicer Dicer is a self-service reporting tool that allows clinicians to customize searches on large patient populations using powerful data exploration tools. Using Slicer Dicer, clinicians can find the data they need to investigate a hunch, and then refine or reformulate their searches on the fly to better understand their patient populations. With access to all the clinical data documented in the chart, physicians are also able to examine trends in their patients. Slicer Dicer can often provide rough counts on the number of patients who meet simple inclusion and exclusion criteria based on demographics, diagnoses, and lab data from EPIC.

- *Accessing Slicer Dicer:*
    - Slicer Dicer is designed to be a starting point for a question you might have and want to explore further. It is not meant to replace any existing reporting tools, but instead to supplement your ability to quickly explore new ideas and decide on what to research further using other reporting tools. It is not meant for proving causality or providing direct care.
    - Slicer Dicer can be accessed via the EPIC button>Reports>Slicer Dicer
- *Use of Slicer Dicer:*
    - Use of slicer dicer to explore preliminary hypotheses and generate estimated counts for funding proposals does not require IRB approval. Slicer Dicer should *not* be used for research study analysis.

❖ See Appendix A for the transcripts of the interviews. See appendix for FAQ on Data Trust review of research data requests, a structured diagram on how to request data from CCDA and use Slicer Dicer (EPIC tool) for preliminary searches

- ***Challenges to Using EPIC Data for BSSS Research***

Challenges to using EPIC data for BSSS research include lack of data harmonization, including heterogeneity of collecting, entering, and visualing data (on the receiving end) across clinics. There is a need for an awareness of existing data, resources, variables, and nuances of variables being collected across JHM which may vary by clinic and department. A single variable such as gender can be collected in 13 different ways, which then impacts how the data is extracted from EPIC and used and interpreted for research purposes. EPIC data collection fields and content may differ clinic by clinic. Local content can be built into EPIC given the potential to have specialized forms with more detailed questions pertinent to a specific clinic. While this capability allows for data collection fields to be tailored to a specific clinic and patient population, it creates difficulty in obtaining data that is stored in different places and results in fragmented. Data may be missing, and we may not realize this on data extraction or on attempts to harmonize the data. Challenges with data collection include clinic-by-clinic variation in who collects the data (i.e., Patient service coordinator on registration, clinical assistant, nurse or physician during visit). Also, there may be differences if data are patient reported or not. For example, race or ethnicity may differ depending on patient reported race/ethnicity, versus perceived race/ethnicity assigned by the data collector.

Regarding missing data, there are some data elements that must be entered, for example, patient name. So 100% of patients should have a name (although it might not be the right name). There are some data elements that had to be entered once EPIC went live (like Race), but might be missing for historical data that was loaded for patients that haven't visited Hopkins again since 2013. Then there are some data elements that are only collected in certain locations (like certain data only collected during an inpatient stay) or data elements only collected by a certain patient population (PSA for men) or by a certain practice (ophthalmology data).

Challenges to EPIC data use also include confounding, bias, handling of missing data, data management, changes in data over time, outliers, patient identification number may not be unique or reliable, especially when merging different data sources into EPIC.

Challenges identified by researchers using MyChart to collect social determinant measures include getting people onto MyChart and identifying workflows that do not burden the staff in the process. Workflow issues were noted to be as important as the technical challenges. Having a simplified workflow system is critical.

Other challenges to consider when using specific BSSS data from EPIC:

- Death data: unless the patient died at a JHM facility or a family member contacts JHM, we don't know for sure if the patient has died.
- Smoking status: the collection accuracy varies from clinic to clinic. Sometimes this question isn't asked.
- Race is captured for most patients (about 4.5 million of the 5.1 million in EPIC).
- Education status is not well captured at the time of admission.
- The absence of a data element doesn't always imply that a behavior wasn't observed – it just may mean that no one asked the question.
- Flowsheets, questionnaires, SmartData can be different across sites. For example, one flowsheet in the ED at JHH could look slightly different (capture different data elements) than a flowsheet in the ED at Sibley.
- Data extracted out of the backend database doesn't always look as well structured as it does in the front-end. The front-end often performs calculations on data (lab values) or makes workflow decisions that don't show up in the database.

- Unstructured notes (pathology notes, radiology notes, progress notes) are not easy to search (although there are many improvements coming that may make this process easier – Natural Language Processing, full text searching).

See Appendix A for the transcripts of the interviews.

- ***Analyzing Existing EPIC Data***

  - *Data Specifications Matrix*

The guide project team, in collaboration with the CCDA, developed a data matrix which defines a series of data dimensions. The data matrix will be used in phase #2 as a coding schema to capture various specs of the social/behavioral data found in the EPIC and other potential data sources. For example, if 'education' is identified as a potentially high impact social variable based on the literature review and then located/found in the EPIC, then a series of specifications about that data element (i.e., education) will be captured/created such as where exactly this data source is shown (on screen) and stored (in database), what are the potential data quality issues (e.g., completeness, accuracy, and timeliness), and what are various data governance issues that may hinder accessing the data by researchers. Followed is an outline of the data matrix that was designed in phase 1 and applied to a select list of social/behavioral data types (see Appendix C):

- What: variable of interest
- Whose: variable exists for this patient denominator
- When: temporal aspects of the variable
- Where: location that the variable is often collected
- Who: person collecting the variable
- Data Management
  - Data provenance (source)
  - Data type
  - Data quality (accuracy, completeness, timeliness)

❖ See Appendix B for the detailed information about the data matrix.

  - *List of JHMI EPIC Social & Behavioral Variables*

We obtained a list of highly recommended variables in the literature and commonly requested from CCDA and applied the human and data matrices. These variables included: race, ethnicity, alcohol use, depression, tobacco use, and residence zip code.

The data matrix was applied against the highly requested variables and the results are available in the appendix.

In Phase 2 we will use the same approach and apply the human and data matrix to additional social/behavioral variables. We will:

- Apply the Human Matrix to new social/behavioral variables
- Apply the Data Specification Matrix to new social/behavioral variables
- List of EPIC tools/instruments/surveys used to collect new social/behavioral

❖ See Appendix B for the data matrix applied to common social/behavioral data.

○ *Data Quality Queries*

In Phase 2, CCDA will run in a query to establish the availability and quality of the social and behavioral variables collected in EPIC. Data completeness will be examined. These results will feed into the data matrix as well.

Table 2 - Valuable and recommended social and behavioral variables that will potentially be further explored in phase 2

NAM Recommended Core Domain	NAM Recommended Measure	Does JH EPIC currently collect in any from?	How is it collected?
Sociodemographic			
Race/ethnicity	US Census	TBD	TBD
Education	Educational attainment	TBD	TBD
Financial Resource Strain	Overall financial resource strain	TBD	TBD
Health Literacy		TBD	TBD
Psychological			
Stress	Elo et al. (2003)	TBD	TBD
Depression	PHQ-2	TBD	TBD
Exposure to violence; Intimate partner violence	HARK	TBD	TBD
Behavioral			
Physical activity	Exercise Vital Sign	TBD	TBD
Tobacco use and exposure	NHIS	TBD	TBD
Alcohol use	AUDIT-C	TBD	TBD
Individual-level Social Relationships			
Social connections and social	NHANES III	TBD	TBD

isolation			
Exposure to violence; Intimate partner violence	HARK	TBD	TBD
Neighborhoods and Communities			
Neighborhood and community compositional characteristics	Residential address	TBD	TBD
	Census tract-median income	TBD	TBD

NAM: National Academy of Medicine

- *Retrieving Social and Behavioral Data from EPIC*

In phase 2, we will complement this report with practical guides on how to request/access the social data. See appendix for a structured diagram on how to request data from CCDA and use Slicer Dicer (an EPIC tool) for preliminary searches. Policies (e.g. data council, IRB, Slicer Dicer) will be covered as they apply to using the variables for clinical care, quality improvement, or research. A guide to explain how external (non-EPIC) data can supplement EPIC data to provide a broader array of social and behavioral data group, MHCC, Medicare, HIE, other providers' EMRs) will be created in Phase 2. We have collated guides to CCDA and Data Trust in the appendix.

- *CCDA's Role and Procedures*

In Phase 2, we will further gather information beyond the guides available in the appendix. We look specifically at the following:

- How to collect new data? (e.g., add new instruments/surveys)
- How to incorporate/integrate external social/behavioral data?
- Individual level (e.g., MHCC, HIE, Medicare/Medicaid)
- Aggregate level (e.g., geo-spatial databases such as Census)
- Language to be used for NIH grants
- List of high-impact social/behavioral variables in EPIC
- Linking external datasets (e.g., trials) with social/behavioral data
- Implication for multi-site studies/trials
- Relevance to "Precision Medicine"
- Methods/technology used to extract/clean social/behavioral data
- HIPAA and IRB implications

❖ See Appendix C for additional details about extracting data from EPIC.

■ DISCUSSION

Overall, the ability to extract social determinant measures from existing databases and medical records is limited by four major factors. First and foremost, most of the measures related to social determinants or their constituent parts are not captured in a systematic fashion in the JHMI EMR. Second, to the extent that measures are available, they have to be constructed/calculated from fields in the databases. Third, the need for database management and research design skills is major shortcoming in many of the requests that are being submitted to CCDA. Lastly, there is no standardized mechanism, protocol, or algorithm for collecting social determinant measures should a researcher wish to conduct a study. Each issue is considered in turn, followed by specific recommendations.

• *Current Social Determinant Data Collection*

Social determinant measures are not strictly speaking necessary to making a medical diagnosis. Moreover, most measures are not an essential element for documenting care and / or receiving reimbursement. Therefore, most measures that would be considered an assessment of a patient's social determinants of health are not documented in a structured field. Nevertheless, it is likely that many clinicians discuss a patient's personal and environmental backgrounds as part of an encounter.

Social determinant factors may be captured in the 'open notes' component of the patient's medical record. Structured fields for social determinant measures could be added to the EMR. However, clinicians are already overburdened with documentation requirements and are likely to resist any additional data collection that does not have a clear medical necessity. Managers are also likely to resist the addition of any measures that extend clinical encounters, require additional information technology or lack reimbursement implications (either negative or positive). Therefore, some other means for capturing social determinants is needed.

• *Calculating and Constructing Social Determinant Measures*

Merging existing patient data from structured fields with other information sources to create new variables may generate valuable social determinant measures. Environmental social determinants (e.g., access to transportation and employment) can be created based on patient's residence in combination with other data sources. Other measures related to socioeconomic status (e.g., income) could also be inferred based on residence, insurance mechanism and other variables that are likely to be captured in the EMR's structured data fields. Variables related to individuals' living arrangements and family histories could be created if EMR records were linked across patients. The latter

set of measures would also have benefits related to checking the accuracy of fields such as race and ethnicity. For example, if an individual's parents have records in the EMR system, measures such as race could be cross-checked with other family members' records. Any discrepancies detected would require a human assessment to reconcile.

One possible source for reconciling discordant data fields and adding information about social determinants is the patient. The PHR is currently being used to collect self-reported data related to social determinants for some research. Each study's protocol and data collection are idiosyncratic to that study. Therefore, the data tends to have limited utility beyond its specific purpose. However, having the patient self-report measures related to their social determinants has many appealing features.

Another existing information source is the 'unstructured' clinical notes contained in the EMR. It may be possible for researchers to mine these notes for social determinant measures using natural language processing and other machine learning algorithms. The use of artificial intelligence for health services research is in its early days and it is unlikely that researchers will have access to such tools in the near-term and must find other means to collect social determinant measures.

- ***Population and Community Health Applications***

Population health management is increasingly becoming an integral part of value-based provider operations. Effective population health management needs reliable risk stratification to better identify patients at high-risk for undesired outcomes.

Although risk stratification has been traditionally developed using administrative claims, EMR data are becoming instrumental for risk stratification among providers [65]. Multiple studies have shown the added-value of EMR data for risk stratification and population health management efforts [66-71]. One of the potential added-values of EMRs for risk stratification is incorporating EMR-derived social determinant factors [72]; however, extracting social factors from EMRs may require dealing with multiple issues such as: EMR maturation [73], data quality issues [74], lack of advanced methods to extract social determinants from EMR's free-text [75], and incorporating additional questionnaires within the EMR's architecture [76,77].

Given the increased role of providers in their communities, population and public health efforts are becoming more aligned [78-81]. Identifying social determinant factors for all patients of a provider network will be a critical element in aligning efforts to address disparities within a provider's catchment area and increase the health of the surrounding communities (specially under Maryland's all-payer waiver program) [82-83]. Non-EMR data sources, such as health information exchange data, can also be used to extract social determinant data [84].

- ***Researcher Competency Enhancement***

There are two main challenges with respect to social determinants' studies arising from research design competencies. The first limitation is researchers' lack of understanding with respect to how EMR data is collected, stored, and extracted for analysis. While most clinical staff members interact with the EMR, the expectation that the fields they see in daily use can be pulled from across the health system or the broader community is mistaken. The same clinical variable may be stored in a variety of fields under different names depending on how the EMR 'build' was undertaken. The magnitude of this issue grows as more organizations or sub-units are added to the requested data pull.

Another common problem with data requests revolves around the identification of populations or patient panels. Many clinicians ask for a panel of subjects with a disease state or set of characteristics with the intention of proposing an intervention. Similar to the identification of specific variables, the variations in data labeling and collection make this task challenging for the data-warehouse without clearer guidance from the researcher. The process of 'walking' a researcher through the data fulfillment task generally proves to be prohibitively expensive and takes too long to meet the researcher's needs. At one point, the I2B2 system was intended to mitigate this issue by providing researchers a simple means for assessing if there was a sufficient population to conduct the envisioned research. However, the system did not effectively meet this aim and the aforementioned "Slicer Dicer" is not yet available. Even when that tool is made available it will not resolve a more fundamental challenge related to research design competencies.

A common refrain across the interviews was that having clearly articulated research hypotheses would greatly help the CCDA serve the customer at-hand. Further still, having a more complete picture of the intended research design would make data collection feasibility questions easier to answer. There are several possible activities and tools that would ameliorate the challenge researchers face in preparing a data request application.

- ***Tools for Facilitating Social Determinants in Research***

Many of the tools that would help researchers develop studies and efficiently request data are topic agnostic.

- ***Current Resources and Next Steps***

Multiple resources at JHM are available to support researchers conducting BSSS research. The BSSS Translational Research Community (TRC) stands at the forefront of

leading and creating a community for researchers from across JHU who are conducting research in the areas of health and behavior, biopsychosocial interactions, social and cultural factors in health, health systems and health services, health IT, and methodologies. Additional resources include the Data Trust Council, Center for Clinical Data Analysis (CCDA), and Institute for Clinical and Translational Research (ICTR).

Current recommendations to guide researchers in using EPIC data for BSSS research includes formulating specific research questions which results in specific requests for data. The Slicer Dicer tool can be used to explore preliminary hypotheses and for more specific data, requests can be submitted to the CCDA.

Next steps and recommendations for facilitation of BSSS research include the development of a web-based flowchart for research, including an interactive step-by-step approach to generating a specific data request. Next steps also include making available a catalog of behavioral and social science-related measures and creating common data collection forms to standardize the collection of social determinant measures from EHR.

In conclusion, while many challenges exist to collecting, extracting, and using EPIC data for BSSS research, community and technical resources are currently available at JHM to support researchers in conducting behavioral, social science, and systems-based research. Further work is needed to continue to improve access to data and the availability of tools to support researchers in conducting BSSS research.

REFERENCES

1. The National Academy of Medicine (NAM) Committee on the Recommended Social and Behavioral Domains and Measures for Electronic Health Records. Capturing Social and Behavioral Domains in Electronic Health Records: Phase 1. Washington (DC); 2014.
2. Ansari Z, Carson NJ, Ackland MJ, Vaughan L, Serraglio A. A public health model of the social determinants of health. Soz Praventivmed. 2003; 48(4):242-51.
3. Feinstein JS. The relationship between socioeconomic status and health: a review of the literature. Milbank Q. 1993; 71(2):279-322.
4. Wen M, Hawkley LC, Cacioppo JT. Objective and perceived neighborhood environment, individual SES and psychosocial factors, and self-rated health: an analysis of older adults in Cook County, Illinois. Soc Sci Med. 2006; 63(10):2575-90.
5. Belanger E, Ahmed T, Vafaei A, Curcio CL, Phillips SP, Zunzunegui MV. Sources of social support associated with health and quality of life: a cross-sectional study among Canadian and Latin American older adults. BMJ Open. 2016; 6(6): e011503.
6. Bosworth HB, Schaie KW. The relationship of social environment, social networks, and health outcomes in the Seattle Longitudinal Study: two analytical approaches. J Gerontol B Psychol Sci Soc Sci. 1997; 52(5):197-205.
7. Rosano A, Loha CA, Falvo R, van der Zee J, Ricciardi W, Guasticchi G, et al. The relationship between avoidable hospitalization and accessibility to primary care: a systematic review. Eur J Public Health. 2013; 23(3):356-60.
8. Salmond C, Crampton P, Sutton F. NZDep91: A New Zealand index of deprivation. Aust N Z J Public Health. 1998; 22(7):835-7.
9. Marmot MG, Smith GD. Why are the Japanese living longer? BMJ. 1989; 299(6715):1547-51.
10. Bandura A. The anatomy of stages of change. Am J Health Promot. 1997; 12(1):8-10.
11. Frenk J. Medical care and health improvement: the critical link. Ann Intern Med. 1998;129(5):419-20.
12. Link BG, Phelan J. Social conditions as fundamental causes of disease. J Health Soc Behav. 1995; Spec No:80-94.
13. Kahn JR, Pearlin LI. Financial strain over the life course and health among older adults. J Health Soc Behav. 2006; 47(1):17-31.
14. Steenland K, Hu S, Walker J. All-cause and cause-specific mortality by socioeconomic status among employed persons in 27 US states, 1984-1997. Am J Public Health. 2004; 94(6):1037-42.
15. Minkler M, Fuller-Thomson E, Guralnik JM. Gradient of disability across the socioeconomic spectrum in the United States. N Engl J Med. 2006; 355(7):695-703.
16. Altman BM, Blackwell DL. Disability in U.S. Households, 2000-2010: Findings from the National Health Interview Survey. Fam Relat. 2016; 63(1):20-38.
17. Spillman BC, Long SK. Does high caregiver stress predict nursing home entry? Inquiry. 2009; 46(2):140-61.
18. Gundersen C, Ziliak JP. Food Insecurity and Health Outcomes. Health Aff (Millwood). 2015; 34(11):1830-9.
19. Bhargava V, Lee JS. Food Insecurity and Health Care Utilization Among Older Adults. J Appl Gerontol. 2016.
20. Ziliak JP GC, Haist M. The causes, consequences, and future of senior hunger in America. 71 ed. Lexington, KY: UK Center for Poverty Research, University of Kentucky; 2008.

21. Berkowitz SA, Seligman HK, Choudhry NK. Treat or eat: food insecurity, cost-related medication underuse, and unmet needs. Am J Med. 2014; 127(4):303-10 e3.

22. Seligman HK, Davis TC, Schillinger D, Wolf MS. Food insecurity is associated with hypoglycemia and poor diabetes self-management in a low-income sample with diabetes. J Health Care Poor Underserved. 2010; 21(4):1227-33.

23. Seligman HK, Laraia BA, Kushel MB. Food insecurity is associated with chronic disease among low-income NHANES participants. J Nutr. 2010 ;140(2):304-10.

24. Vozoris NT, Tarasuk VS. Household food insufficiency is associated with poorer health. J Nutr. 2003; 133(1):120-6.

25. Winkleby MA, Jatulis DE, Frank E, Fortmann SP. Socioeconomic status and health: how education, income, and occupation contribute to risk factors for cardiovascular disease. Am J Public Health. 1992; 82(6):816-20.

26. Mensah GA, Mokdad AH, Ford ES, Greenlund KJ, Croft JB. State of disparities in cardiovascular health in the United States. Circulation. 2005 ;111(10):1233-41.

27. Freedman VA, Spillman BC. Active Life Expectancy in The Older US Population, 1982-2011: Differences Between Blacks And Whites Persisted. Health Aff (Millwood). 2016; 35(8):1351-8.

28. Maddox TM, Reid KJ, Spertus JA, Mittleman M, Krumholz HM, Parashar S, et al. Angina at 1 year after myocardial infarction: prevalence and associated findings. Arch Intern Med. 2008; 168(12):1310-6.

29. Weaver WD, White HD, Wilcox RG, Aylward PE, Morris D, Guerci A, et al. Comparisons of characteristics and outcomes among women and men with acute myocardial infarction treated with thrombolytic therapy. GUSTO-I investigators. JAMA. 1996; 275(10):777-82.

30. Zusterzeel R, Selzman KA, Sanders WE, Canos DA, O'Callaghan KM, Carpenter JL, et al. Cardiac resynchronization therapy in women: US Food and Drug Administration meta-analysis of patient-level data. JAMA Intern Med. 2014;174(8):1340-8.

31. Nicholson A, Kuper H, Hemingway H. Depression as an aetiologic and prognostic factor in coronary heart disease: a meta-analysis of 6362 events among 146 538 participants in 54 observational studies. Eur Heart J. 2006; 27(23):2763-74.

32. Dong JY, Zhang YH, Tong J, Qin LQ. Depression and risk of stroke: a meta-analysis of prospective studies. Stroke. 2012; 43(1):32-7.

33. Pinquart M, Duberstein PR. Depression and cancer mortality: a meta-analysis. Psychol Med. 2010; 40(11):1797-810.

34. Reynolds SL, Haley WE, Kozlenko N. The impact of depressive symptoms and chronic diseases on active life expectancy in older Americans. Am J Geriatr Psychiatry. 2008; 16(5):425-32.

35. Ferrari AJ, Charlson FJ, Norman RE, Patten SB, Freedman G, Murray CJ, et al. Burden of depressive disorders by country, sex, age, and year: findings from the global burden of disease study 2010. PLoS Med. 2013; 10(11): e1001547.

36. Pearlin LI. The sociological study of stress. J Health Soc Behav. 1989; 30(3):241-56.

37. Adler NE, Stewart J. Health disparities across the lifespan: meaning, methods, and mechanisms. Ann N Y Acad Sci. 2010; 1186:5-23.

38. Sandel M, Wright RJ. When home is where the stress is: expanding the dimensions of housing that influence asthma morbidity. Arch Dis Child. 2006; 91(11):942-8.

39. Fagerstrom K. The epidemiology of smoking: health consequences and benefits of cessation. Drugs. 2002; 62 Suppl 2:1-9.

40. McKnight-Eily LR, Liu Y, Brewer RD, Kanny D, Lu H, Denny CH, et al. Vital signs: communication between health professionals and their patients about alcohol use--44 states and the District of Columbia, 2011. MMWR Morb Mortal Wkly Rep. 2014; 63(1):16-22.

41. Greene J, Hibbard JH. Why does patient activation matter? An examination of the relationships between patient activation and health-related outcomes. J Gen Intern Med. 2012; 27(5):520-6.

42. Greene J, Hibbard JH, Sacks R, Overton V, Parrotta CD. When patient activation levels change, health outcomes and costs change, too. Health Aff (Millwood). 2015; 34(3):431-7.

43. Mosen DM, Schmittdiel J, Hibbard J, Sobel D, Remmers C, Bellows J. Is patient activation associated with outcomes of care for adults with chronic conditions? J Ambul Care Manage. 2007; 30(1):21-9.

44. Remmers C, Hibbard J, Mosen DM, Wagenfield M, Hoye RE, Jones C. Is patient activation associated with future health outcomes and healthcare utilization among patients with diabetes? J Ambul Care Manage. 2009; 32(4):320-7.

45. Kinney RL, Lemon SC, Person SD, Pagoto SL, Saczynski JS. The association between patient activation and medication adherence, hospitalization, and emergency room utilization in patients with chronic illnesses: a systematic review. Patient Educ Couns. 2015; 98(5):545-52.

46. Begum N, Donald M, Ozolins IZ, Dower J. Hospital admissions, emergency department utilisation and patient activation for self-management among people with diabetes. Diabetes Res Clin Pract. 2011; 93(2):260-7.

47. Hendriks M, Rademakers J. Relationships between patient activation, disease-specific knowledge and health outcomes among people with diabetes; a survey study. BMC Health Serv Res. 2014; 14:393.

48. Skolasky RL, Mackenzie EJ, Riley LH, 3rd, Wegener ST. Psychometric properties of the Patient Activation Measure among individuals presenting for elective lumbar spine surgery. Qual Life Res. 2009; 18(10):1357-66.

49. Graven LJ, Grant JS. Social support and self-care behaviors in individuals with heart failure: an integrative review. Int J Nurs Stud. 2014; 51(2):320-33.

50. Lee KS, Lennie TA, Yoon JY, Wu JR, Moser DK. Living Arrangements Modify the Relationship Between Depressive Symptoms and Self-care in Patients with Heart Failure. J Cardiovasc Nurs. 2016.

51. Mu C, Kecmanovic, M., & Hall, J. Does Living Alone Confer a Higher Risk of Hospitalization. Economic Record. 2015; 91(S1):124-38.

52. Udell JA, Steg PG, Scirica BM, Smith SC, Jr., Ohman EM, Eagle KA, et al. Living alone and cardiovascular risk in outpatients at risk of or with atherothrombosis. Arch Intern Med. 2012; 172(14):1086-95.

53. Redfors P, Isaksen D, Lappas G, Blomstrand C, Rosengren A, Jood K, et al. Living alone predicts mortality in patients with ischemic stroke before 70 years of age: a long-term prospective follow-up study. BMC Neurol. 2016; 16:80.

54. Schmaltz HN, Southern D, Ghali WA, Jelinski SE, Parsons GA, King KM, et al. Living alone, patient sex and mortality after acute myocardial infarction. J Gen Intern Med. 2007; 22(5):572-8.

55. Manzoli L, Villari P, G MP, Boccia A. Marital status and mortality in the elderly: a systematic review and meta-analysis. Soc Sci Med. 2007; 64(1):77-94.

56. Molloy GJ, Stamatakis E, Randall G, Hamer M. Marital status, gender and cardiovascular mortality: behavioural, psychological distress and metabolic explanations. Soc Sci Med. 2009; 69(2):223-8.

57. Schwandt HM, Coresh J, Hindin MJ. Marital Status, Hypertension, Coronary Heart Disease, Diabetes, and Death Among African American Women and Men: Incidence and Prevalence in the

Atherosclerosis Risk in Communities (ARIC) Study Participants. J Fam Issues. 2010; 31(9):1211-29.

58. Duru OK, Vargas RB, Kermah D, Pan D, Norris KC. Health insurance status and hypertension monitoring and control in the United States. Am J Hypertens. 2007; 20(4):348-53.
59. Gandelman G, Aronow WS, Varma R. Prevalence of adequate blood pressure control in self-pay or Medicare patients versus Medicaid or private insurance patients with systemic hypertension followed in a university cardiology or general medicine clinic. Am J Cardiol. 2004; 94(6):815-6.
60. Andersen ND, Brennan JM, Zhao Y, Williams JB, Williams ML, Smith PK, et al. Insurance status is associated with acuity of presentation and outcomes for thoracic aortic operations. Circ Cardiovasc Qual Outcomes. 2014; 7(3):398-406.
61. Gaskin DJ, Thorpe RJ, Jr., McGinty EE, Bower K, Rohde C, Young JH, et al. Disparities in diabetes: the nexus of race, poverty, and place. Am J Public Health. 2014; 104(11):2147-55.
62. Diez-Roux AV, Nieto FJ, Muntaner C, Tyroler HA, Comstock GW, Shahar E, et al. Neighborhood environments and coronary heart disease: a multilevel analysis. Am J Epidemiol. 1997; 146(1):48-63.
63. O'Campo P, Xue X, Wang MC, Caughy M. Neighborhood risk factors for low birthweight in Baltimore: a multilevel analysis. Am J Public Health. 1997; 87(7):1113-8.
64. Sullivan CG. Putting " health" in the electronic health record: A call for collective action. Nursing Outlook. 2015; 63(5):614-6.
65. Kharrazi H, Lasser E, Yasnoff WA, Loonsk J, Advani A, Lehmann H, Chin D, Weiner JP. A proposed national research and development agenda for population health informatics: summary recommendations from a national expert workshop. J Am Med Inform Assoc. 2017; 24 (1):2-12
66. Kharrazi H, Chi W, Chang HY, Richards TM, Gallagher JM, Knudson SM, Weiner JP. Comparing population-based risk-stratification model performance using data extracted from electronic health records versus administrative claims. Med Care. 2017; 55 (8): 789-796
67. Kharrazi H, Weiner JP. A practical comparison between the predictive power of population-based risk stratification models using data from electronic health records versus administrative claims: setting a baseline for future EHR-derived risk stratification models. Med Care, 2017; 56(2), 202-203
68. Chang HY, Richards TM, Shermock KM, Elder-Dalpoas S, Kan H, Alexander CG, Weiner JP, Kharrazi H. Evaluating the impact of prescription fill rates on risk stratification model performance. Med Care. 2017; 55 (12): 1052-1060
69. Kan H, Kharrazi H, Leff B, Boyd C, Davison A, Chang H-Y, Kimura J, Wu S, Anzaldi LJ, Richards T, Lasser E, Weiner JP. Defining and assessing geriatric risk and associated health care utilization among elderly patients using claims and electronic health records. Med Care. 2018; 56(3): 233-239
70. Lemke K, Gudzune KA, Kharrazi H, Weiner JP. Assessing markers from ambulatory laboratory tests for predicting high-risk patients. Am J Manag Care. 2018; 24(6): e190-e195
71. Kharrazi H, Chang HY, Heins S, Weiner JP, Gudzune K. Enhancing the prediction of healthcare costs and utilization by including outpatient BMI values to diagnosis-based risk models. Med Care. 2018; 56 (12): 1042-1050
72. Hatef E, Searle KM, Predmore Z, Lasser EC, Kharrazi H, Nelson K, Sylling P, Curtis I, Fihn S, Weiner JP. The impact of social determinants of health on hospitalization in the Veterans Health Administration. Am J of Prev Med. In-press.
73. Kharrazi H, Gonzalez CP, Lowe KB, Huerta TR, Ford EW. Forecasting the maturation of electronic health record functions among US hospitals: retrospective analysis and predictive model. J Med Internet Res. 2018; 20(8): e10458

74. Kharrazi H, Wang C, Scharfstein D. Prospective EHR-based clinical trials: the challenge of missing data. J Gen Intern Med. 2014; 29 (7): 976-978

75. Kharrazi H, Anzaldi L, Hernandez L, Davison A, Boyd CM, Leff B, Kimura J, Weiner JP. Measuring the value of electronic health record's free text in identification of geriatric syndromes. J Am Geriatr Soc. 2018; 66(1) 1499-1507

76. Wu A, Kharrazi H, Boulware LE, Snyder CF. Measure once, cut twice – adding patient reported outcome measures to the electronic health record for comparative effectiveness research. J Clin Epidemiol. 2013; 66 (8): S12-20

77. Bae J, Ford EW, Kharrazi H, Huerta TR. Electronic medical record reminders and smoking cessation activities in primary care. Addict Behav. 2017; 16 (77): 203-209

78. Kharrazi H, Weiner JP. IT-enabled community health interventions: challenges, opportunities, and future directions. Generating Evidence & Methods to Improve Patient Outcomes (eGEMs). 2014; 2 (3): 1-9

79. Dixon B, Kharrazi H, Lehman H. Public health and epidemiology informatics: recent research and events. Yearb Med Inform. 2015; 10 (1): 199-206

80. Dixon B, Pina J, Kharrazi H, Gharghabi F, Richards J. What's past is prologue: a scoping review of recent public and global health informatics literature. Online J Public Health Inform. 2015; 7 (2) e1-31

81. Gamache R, Kharrazi H, Weiner JP. Public health and population health informatics: the bridging of big data to benefit communities. Yearb Med Inform. 2018; 27(1): 199-206

82. Hatef E, Kharrazi H, VanBaak E, Falcone M, Ferris L, Mertz K, Perman C, Bauman A, Lasser EC, Weiner JP. A state-wide health IT infrastructure for population health: building a community-wide electronic platform for Maryland's all-payer global budget. Online J Public Health Inform. 2017; 9(3): e195

83. Hatef E, Lasser EC, Kharrazi H, Perman C, Montgomery R, Weiner JP. A population health measurement framework: evidence-based metrics for assessing community-level population health in the global budget context. Popul Health Manag. 2017; 21(4): 261-270

84. Kharrazi H, Horrocks D, Weiner JP. Use of HIEs for value-based care delivery: a case study of Maryland's HIE. In Dixon B (Ed.) Health Information Exchange: Navigating and Managing a Network of Health Information Systems. 2016; 313-332. Cambridge, MA: Academic Press Elsevier

■ APPENDIX A – INTERVIEW NOTES/TRANSCRIPTS

• *Semi-Structured Interview with D. Gumas*

- Raw vs transformed data
    - Diana Gumas – emphasized her perspective as a programmer
    - Diana – gets data in raw form
    - Many other departments transform the data
    - Jenny Bailey – would be good person to interview
    - Derived – set of data – perhaps
    - In the quality improvement work, might she be deriving some things that are social determinants
- Need for greater awareness of existing data, resources, variables, and nuances of variables being collected across JHM - departments/clinics
    - What are people collecting other than the standard variables?
    - Brandon Lau –collecting gender in 13 different ways.
    - Work with clinical colleagues – build items
    - Albert Wu – runs questionnaire committee – patient reported outcomes
    - Physician – standard workflow – specialized tweaking in each setting
    - Feature in EPIC to share?
- Challenge: What is the local content that we built?
    - Not the same across the board. Specialized forms with more detailed questions on pertinent information to a specific clinic – i.e. HIV clinic – want to know more nuance about info in a certain clinic - ask specialized questions about sexual activity – then ask about broken bone, then ask about more questions of specific interest.
    - From a clinician's point of view –data in multiple places – hard to find or reconcile (if same question answered differently in 2 different places)
- Challenge: Data Harmonization
    - Data harmonization is part of precision medicine platform, led by Chris Chute
    - Some efforts on harmonization of data in the warehouse – just learning how to do this
    - Fragmented data – data missing and we don't even know it.
    - How much uniformity do we want and how much value is there in variation?
- Challenges: Data Collection

- different from each clinic – different role collects in different clinics
- Patient reported vs data collector assume (i.e., race/ethnicity)
- EPIC programmers
- Program view – lots (JK: not sure what this refers to)

- Challenge: IT Human Resources (noted below)
- Types of Data Requests
  - A distinction between two types of data requests: (a) building data collection into EPIC; and, (b) getting data out of EPIC
    - (a) <u>Building data collection into EPIC</u>
      - Diana runs EPIC research team – ordersets, research building, maintenance – 3 member team
      - Just last week got enhancements to build for research
      - Build me a specialized view
      - Just getting to that now
    - (b) <u>Getting data out of EPIC (for research)</u>
      - More mature processes to address this. Five people are trained to this. A year and half ago, it took 1-2 weeks to respond to a request, now much faster turn around time.
      - A year ago, the data trust process took at long time and was an impediment to obtaining data for research. Now, they are only reviewing request if identifiable data is going out of Hopkins or for requests involving many patients (i.e., 10k patients in data set).
      - Now if a study is IRB approved for 400 pts and it is conducted at Hopkins on secure server, then data trust does not come into play.
      - Process has become streamlined so that ICTR can respond rapidly with fewer bumps in the road.

- Follow up questions for Diana Gumas
  - What are the first steps that you would recommend to someone looking to OBTAIN DATA from EPIC?
  - What are the first steps that you would recommend to someone looking to BUILD DATA in EPIC?
  - Please provide examples of well-structured requests for data
  - List of most common data queries to include in the guide – with estimates of cost
  - Catalog of existing data (Chris Chute)
  - Data dictionary – explanation and quality of variables (Chris Chute)

    - Slicer Dicer PDF handouts
    - Organizational chart of data - how ICTR and CCDA fits into data trust council org chart?
    - List of 10 centers – Johns Hopkins Data Trust
- Additional Resources
    - Slicer Dicer
        - went live in January
        - Available to 26,000 people (if EPIC access, see patients, on IRB approved research study, ?medical students) - currently does not have anyone
        - Challenge: non-clinician researcher getting access to SlicerDicer (If JHSPH was part of covered entity, then would address these challenges, but at this time, they are not).
    - ICTR
        - 2 free hours for service – how does that work? See website – enter info.
- Various Issues
    - EPIC Builds (building patient reported outcomes fields, decision support, etc)
    - Entering data
    - Visualization of data (receiving end)
    - Here are in general the inputs that we are missing
    - EPIC /MEASURE is working on various aspects
    - Need to harmonize across JHM
- Building Data vs Getting Data Out
    - These activities involves two separate teams, two separate approval process. And two separate financial structures.
- Other comments
    - Tableau is a tool for visualizing and exploring data. Can request: visualization of these 25 data elements – yes/no patient identifiable data. What is your ideal thing? Drill down, chart, etc. Can train to build tableau – need to be on. Do not need to go to EPIC for this.
    - Center for clinical data analysis (CCDA): Diana runs this group, Bonnie Woods – is the manager of this. CCDA is one of 10 analytic teams that reports up to the data trust. Currently only 1 person from each 10 analytic team can build tableau.
    - Data layer bringing together values from EPIC system – simpler to learn, build by 10 analytic groups. Do not have to go to EPIC to use tableau.

        - Build tableau unit, building on work of EPIC – leveraging work already done in data layer – tagged as social determinants
        - We may want to create another class of users who are tableau trained (much lower cost) focused to produce visualizing and tables, vs SQL ($10,000, 4 months to barely be able to do this) where you are learning to program.
    - Recommendation: Add an adjunct programmer to population health department.

- ***Semi-Structured Interview with D. Gumas and B. Woods***

1. **What are the first steps that you would recommend to someone looking to OBTAIN DATA from EPIC?**

    - *They should think carefully about what data are needed. I recommend outlining it as follows:*
    - *For what patients do you desire the data? (e.g. all patients for which I am the PCP, or all patients who meet a set of inclusion and exclusion criteria approved by the IRB, or all patients consented to my study and actively on study in the Clinical Research Management System.*
    - *For what time frame to do you desire the data?*
    - *From what locations do you desire the data? (e.g. Johns Hopkins Hospital? Bayview Medical Center? Johns Hopkins Community Physicians? Sibley Memorial? Suburban Hospital? Howard County General? All of the above?)*
    - *Which data elements do you desire? (e.g. race and ethnicity, year of birth, smoking status, diagnoses, etc.). It helps a great deal to partner with a physician who actively uses EPIC who can help you take screen shots of data elements that are more unusual.*
    - *I then recommend contacting the CCDA to ask for an estimate of the cost for a programmer to extract these data for you so that you can then seek funding if needed.*

2. **What are the first steps that you would recommend to someone looking to BUILD DATA in EPIC?**

    - *I am assuming by this question you mean to collect new data elements in EPIC that are not currently collected. If so, then the first step is to meet with the Department/Division/Clinic that you would expect to be collecting these data to get their guidance and buy-in on who should enter the data (the nurse? the physician? the patient? the registrar?) and how that data should be collected. For example, if in the clinical workflow then where that fits into the clinical workflow (a new field on an existing form? a new data collection form?). If being collected from the patient, then is this via MyChart? Or in clinic via the welcome kiosk or on a tablet? Then the request (with support from the affected clinicians who would have to collect the data) will need to be taken to the appropriate Johns Hopkins EPIC committee for consideration. The following link provides info about how to do that. Note that you may have to use VPN to see this page. I couldn't get to it from guest net at Hampton House.*

https://cscop.jhmi.edu/confluence/display/EPIC/Enhancement+Request+Management

3. Please provide examples of well-structured requests for data (Bonnie)

- *Example 1: Adult patients (ages >= 18) seen as outpatients at Bayview and JHH psychiatric clinics from October 1, 2016 to April 30, 2017 diagnosed with major depressive disorder, bipolar disorder, or schizophrenia (either as an encounter diagnosis or on the problem list) having a smoking status that is not "Never". (answers the question "which patient", what encounter type (outpatient vs. inpatient), what encounter location (specific Bayview and JHH psychiatric clinics), what time frame, and other criteria (diagnoses and smoking status).*
- *Example 2: All patients with an in-person (outpatient) visit to a Johns Hopkins internal medicine, family medicine, pediatric, psychiatric, pediatric psychiatric or obstetrics/gynecology clinic from April 1, 2013 until July 1, 2016 whose clinician completed the depression screening flowsheet during that visit. See Appendix A for complete list of departments to include.*

4. List of most common data queries to include in the guide – with estimates of cost. (Bonnie)

- *This is very difficult to provide. In fact, I am working with my staff on a list of common requests and estimates that can be applied to each request (e.g., one database to query with two or three criteria = x hours; two databases to join to match identity and then extract labs and diagnoses = x hours; flowsheet data = x hours; note parsing/searching = x). I'm hesitant to publish anything to researchers right now for fear that they will interpret it as policy.*
- *Very few extracts can be completed under 8-10 hours – I am comfortable in saying this (and do say it on intake calls). The 2 hour complimentary service is usually spent determining requirements, writing spec documents, reviewing requirements with the researcher, and providing an estimate. It's more costly to request data from multiple databases for wide time ranges, and it's more costly to request flowsheet data, questionnaire data, and SmartData, especially without a screen shot or help of a clinician to identify where on the front end the data is presented. Our largest project was 330 hours; the average project is about 30-35 hours.*

5. Catalog of existing data (Chris Chute)

- *A noble goal, but a VERY complex answer that people go to training for weeks to learn and then have to look up a data schema that is many pages long. I think we could give a high level listing of data elements like the following if it would be useful. Please take a look and let me know if this would be of any use at all.*
- *Types of data: Demographics; Encounters - inpatient & outpatient; Vital Signs - e.g. height, weight, blood pressure; Labs; Medications; Diagnoses; Images; Text results; Clinician entered text notes; Patient Questionnaires; Practice-specific data collection forms; Other flowsheet data besides vitals, which may contain patient-reported pain ratings, comfort level/mobility, etc. If this level of detail is useful let us know and Bonnie could make a list of the primary categories*

6. **Data dictionary – explanation and quality of variables (Chris Chute)**

- *This does not exist today except in people's heads. It is something that might either eventually be championed by Chris Chute and the CTSA informatics core and/or the Precision Medicine initiative.*

7. **Organizational chart of data systems – how do ICTR and CCDA fit into the data trust council org chart?**

- *On the following page, the CCDA is one of the analytic teams in the blue box that says Enterprise Analytic Teams*

  http://intranet.insidehopkinsmedicine.org/data_trust/data-trust-organization/

8. **Is there boilerplate language that can be provided to the researcher about EPIC data limitations?**

- *I did write something at some point about the limitations on when we started collecting data at different institutions. Bonnie might have that. If not, let me know and I'll see if I can find it.*
- *I have a chart of when different data elements were backfilled into EPIC and for what categories of data (see attached), as well as a great slide that Diana also put together on how to structure data requests. I also have a few quick limitations that I can think of here:*

  - *Death data (unless the patient died at a JHM facility or a family member contacts JHM, we don't know for sure if the patient has died.*
  - *Smoking status – collection accuracy varies from clinic to clinic. Sometimes this question isn't asked.*
  - *Race is captured for most patients (about 4.5 million of the 5.1 million in EPIC).*
  - *Education status is not well captured at the time of admission.*
  - *The absence of a data element doesn't always imply that a behavior wasn't observed – it just may mean that no one asked the question.*
  - *Flowsheets, questionnaires, SmartData can be different across sites. For example, one flowsheet in the ED at JHH could look slightly different (capture different data elements) than a flowsheet in the ED at Sibley.*
  - *Data extracted out of the backend database doesn't always look as well structured as it does in the front-end. The front-end often performs calculations on data (lab values) or makes workflow decisions that don't show up in the database.*
  - *Unstructured notes (pathology notes, radiology notes, progress notes) are not easy to search (although there are many improvements coming that may make this process easier – Natural Language Processing, full text searching).*

- *I guess my most common caveat that I mention in intake meetings is that clinical data is only as reliable as the clinicians and coders entering the data. "Garbage in, garbage out"*

9. Can we use EPIC data to evaluate gaps in the data, or create a model to predict correct assignment of variables?

- *You could use EPIC to evaluate gaps in data. One simple way to do that, for some data elements like race, would be to use SlicerDicer to find how many patients have an assigned race. Not sure what is meant by a model to predict correct assignment of variables. One thing we did when we set up the EPIC data warehouse was write some queries to look for obviously wrong data, like patients 2 inches high or weighing 2000 pounds. A CCDA data analyst or adjunct member could write queries like that. I have no idea how you could predict correct assignment of something like race.*

10. How does a researcher best address missing data in EPIC?

- *Is the question how to identify that data are missing? Or fix data collection mechanisms so that prospectively data are better collected? Or fix missing data retrospectively?*

11. What % discrepancy in data is due to data variability and issues of health disparity?

- *No idea. Good idea for a research study.*

12. Looking at these data across patients – what % are missing? From what departments? Is there a difference between data quality from ED/Inpatient/and outpatient settings?

- *It really depends on the data element. There are some data elements that have to be entered, for example, patient name. So 100% of patients should have a name (it might not be the right name). There are some data elements that had to be entered once we went live with EPIC (like Race) but might be missing for historical data that was loaded for patients that haven't visited Hopkins again since 2013. Then there are some data elements that are only collected in certain locations (like certain data only collected during an inpatient stay) or data elements only collected by a certain patient population (PSA for men) or by a certain practice (opthmalogy data)*

13. How do you deal with EPIC data with different sources of response options? And, how does this impact how I analyze and interpret the data? What are the response options for these variables? i.e. Free text, options available to choose from, (i.e. Some data sources only have white/black/other options for race, other sources have more options, etc.)

- *We would need to have a conversation about this question. Too complex to put in an email.*

- ***Semi-Structured Interview with V. Smothers***

Responsibilities of the Data Trust

- Leverages EPIC Registries
    - EPIC can take cohort of specific disease, and create registries that they follow
    - Create a registry of patients that meets all the criteria which facilitates all the analytics
- Quality related efforts related to this work
- How to secure and merge data collected across institutions in a place

How to Obtain Information on Data Trust

- Intranet inside Hopkins.org
  http://intranet.insidehopkinsmedicine.org/data_trust/
- Requesting data through data trust:
  http://intranet.insidehopkinsmedicine.org/data_trust/requesting_data_from_an_analytic_team.html
- CCDA consulting group:
  https://johnshopkins.corefacilities.org/service_center/show_external/3796
- Website on Data Trust on Inside Hopkins Medicine
- Link for general FAQ, within that is research-specific FAQ:
  http://intranet.insidehopkinsmedicine.org/data_trust/research-data-requests.html

Typical Reasons for Researchers Go Through the Data Trust

- Sharing data with another institution has to go through the data trust
- Going through another school at Hopkins, like School of Engineering
- Outside of the covered entity includes to the School of Public Health, School of Engineering
- Schools use the Mount Washington data center
- Specific legal counsel on this: within the HIPAA office Pamela Rain mainly with business-associated agreements, Theresa Colescia who is university council focused on research

Organization of Data Trust Council

- Oversight body for data governance in the institution
- That's data in any of our clinical systems, billing systems, the case mix
- Reason why: Now that we have all this data from 5 hospitals, we need centralized oversight, so it provides that
- Data Trust Council has a research specific section that reviews research projects, big projects requests a certain amount of data, often IRB flags it and sends it for review

- ORA sometimes flags things for Data Trust Council review, sometimes researchers themselves ask for review to make sure they were using best review
- There is a quality-specific council that
- Data stewardship council that is looking at how are we taking care of our data, how are we securing it? How are we storing it so people can access it and use it?
- Goes of Data Trust is to coordinate efforts across the institution and reduce redundant effort
- Teams are responsible for analytic work across the institution
- See Figure App A1 for further information about the organizational chart

Figure App A1 – Organizational chart of the Johns Hopkins Data Trust Council

- ***Semi-Structured Interview with D. Thiemann and B. Woods***

Question: Please describe 2 to 3 large gaps in that researchers should be aware of while making requests for data extraction from EPIC.

1. Assumption that EPIC data is clear – it is not. It is "like sipping from a very dirty water hose."
    a. Variable completion rates
    b. Generally systematic biased
    c. For example, if 3/5 elements not filled
    d. Missing data has meaning

2. Most people coming through door do not have any idea about how enterprise data works, or what is in them.
    a. Legacy system database, UB90
    b. From 2012, need to go to completely different system

3. Basics of epidemiology
    a. Many times, it feels like the process involves giving an epi 101 review on "Designing Clinical Research" to assist with the researcher defining their research question and hypothesis.

4. They try to narrow the door to art of possible
    a. Completion rates
    b. Helping to hone queries vs shotgun approach

5. Interface between clinical EMR and research is messy
    a. Rating scale revised 5x in 3 year period
    b. Data retrieval and analysis is similar to archeology
    c. Fall scale morphed and renamed 3x, or changes in required variables / drop down menus – these changes affect query and how scientifically approach
    d. Myth that the data are monolithic and stable – it is constantly evolving
    e. Labs change range of normal
    f. Labs reported in 4 formats (WBC vs WBCx)
    g. Departments come and go
    h. EMR – what maps to what - "the stinking yellow trail"
    i. False notion that EMR research is quick or easy

6. Recommendations to researchers requesting data from EPIC:

    a. Refer to book on designing clinical research: Hulley SB, Cummings SR, Browner, WS, Grady DG, Newman TB. "Designing Clinical Research," 3rd ed. Philadelphia: Lippincott, 2007.
        i. Good users of EHR at Hopkins: Drs. Richard Moore, Graham, Suchisan

    b. Start with a hypothesis, not a content domain, because of data security requirement.
        i. Cannot build your own registry on excel
        ii. Requires more rigorous data management capabilities
            1. Registry about pregnant women with trauma
            2. Cannot just ask for everyone with colorectal surgery – usually not hypothesis driven.

7. Variable specific comments
    a. Smoking: captured
    b. EtOH: [to be completed]
    c. Substance abuse: clinic records (not systems wide data collection), so difficult if not impossible to capture
    d. SES – some pediatricians record, but not consistent documentation
    e. Family support /family history/social history – does not exist in any form that is easily captured. In some clinics it is integrated into flowsheets, but it is not consistently populated. So, if you are looking for info on second hand smoke, data may not reflect a real sampling of patients.

8. Challenges:
    a. Customization of data for every unit, floor, department
    b. Merging of different data elements and forms – difficult to merge
    c. Even with blood pressure reading – there are multiple readings in one visit, which one?
    d. Need to disentangle: smart forms, smart phrases, smart text, free text
        i. Natural lapses in software
        ii. Not well tagged as in XML data
        iii. Not as structured
    e. Data issues:
        i. Confounding
        ii. Bias
        iii. Handling of Missing Data
        iv. Data Management – this is a big gap for researchers requesting data
        v. Changes over time
        vi. Outliers
        vii. MRN may not be unique or reliable, especially merging different data sources into EPIC
    f. Data Management
    g. Diagnoses / Case-finding / Defining your patient population is a challenge:
        i. 23% have chronic kidney disease on problem list
        ii. use complex criteria 2 out 3 to define, vs ICD-10 codes
        iii. Finding cases by ICD-10 codes is problematic
            1. Invalid research
            2. Underestimate
        iv. Challenge in proving that the data is accurate – if not done, and then this creates false science.
        v. This is more so in the outpatient setting, where your search based on a single diagnosis. Less so in inpatient side, because coder abstracts the chart / regulated in Maryland for HSCRC.
        vi. For CKD identified by ICD codes, you would miss 15-40% of patients with that disease.
        vii. There is a need to educate about the limitations with the data.
    h. We do not collect a lot of behavioral and social sciences data in a structured way (pediatrics is somewhat better) – this introduces systemic bias into the data

9. What data is reliable?
    a. Inpatient medications are reliable

10. Can I build data collection into EPIC
    a. yes, you can put a questionnaire in MyChart

11. What if I need preliminary data for my grant?
    a. They can provide basic preliminary data (ie counts or "feasibility" data)
    b. Counts – number of eligible patients - subject to all limitations described above, with very specific eligibility criteria to define your population: i.e. How many patients on medications for the 3 prior visits, were Cr is >x or <y.

12. Three separate divisions in data
    a. Community Hospital Division: Sibley, Suburban, Howard County
    b. Academic Division: 2 academic hospitals
    c. JHCP Division: OP clinics, SOM/JHCP
    d. Many OP clinics have different workflows, did not have EPIC modifications, etc.

13. EPIC backlog
    a. 10 year log
    b. legacy
    c. UB92-data
    d. Casemix / Datamart data
    e. Old EPR 2020, EPM, Casemix, CMRS, direct sequel write
    f. MRN is not unique and reliable!
    g. UGM across institutions – feed data to EPIC, this data is not uniform
    h. Challenge especially for amalgamating social determinants data into EPIC
    i. 20% works with EPIC code, not easy to share across system
    j. Basic data structure may not be the same

14. Costs
    a. Costs increase when you query 2, or 3, or 4 systems
    b. Data is expensive

- ***Semi-Structured Interview with P. Zandi***

- What are you doing? Not yet capturing social determinants. We (NNDC) are capturing patient reported data on mental health and depression as part of a national network (~25 mood disorder clinics). 'Measurement-based care' using a self-reported item.
    - Mania, adverse child experiences.
        - PHQ9, GAD7, 5-items on mania, Columbia suicidal scale (7). Total of 28 items to be completed in the waiting room prior to every visit. Goal is to make it a 'cultural' norm like having their blood pressure taken. In real-time the clinician can see the trended results with potential problems flagged. Thresholds are the trigger.
        - Workflow issues.
            - Questions like, can the survey go out the day before? Decided they wanted it in the waiting room. If they received information outside the clinic, they would have to address them, which might be challenging.

        - Want it in the clinical encounter. The immediate reinforcement increases the notion that it is part of the 'clinical encounter'.
        - Collects the measures through MyChart in the waiting room.
        - The consortium developed a web-based tool for collecting the measures and feeding it back to the clinicians. Therefore, JHMI moved away from MyChart to the consortium tool to create the shared database. The common registry only has the 4 scales. Will eventually move back to EPIC and create web-views, etc. with the clinical data integrated.
        - Next, steps will be to have the richer data with Rx and Dx.
    - New initiative to pull together a team to collect similar tools within the Department of Psychology. CCDA adjunct to work in conjunction with ICTR.
    - People don't know how to approach the ICTR? Worry about being in the queue for data. Building the query tools within the Department (Schizophrenia, Dementia). Patient identification is a big topic.
    - Hoping to get information from the family.
  - How do you define social determinants?
    - Life experiences, SES, race, ethnicity, education.
  - What has been the most difficult challenge in collecting social determinant variables you have faced? No comment
  - What kinds of issues arose? No comment
- Availability of social determinant measure in current existing data collection:
  - Does the EPIC electronic medical record contain the social determinant measures you need for your research?
    - EPIC is building the psychiatry scales back into base system.
    - Psychiatry would like to have: (1) stressful life events; and, (2) much of the important information appears in the notes.
  - Are the data fields routinely filled by patients, administrative staff and other clinical providers? If not, why do you believe they are missing?
- Technical questions:
  - What are the barriers and facilitators to collecting social determinant measures? Simply getting people onto the MyChart is a challenge. Workflows that don't burden the staff in the process. Simplifying the system is critical. Login and passwords are a big issue. Having biometrics would be useful. "*The workflow issues are as important as the technical challenges.*" Have to manually deploy the survey when the patient appears. Creating an automatic trigger.
  - Does the Institute for Clinical and Translational Research (ICTR) provide the necessary training to extract needed social determinant measures? If not, what other opportunities would you like?
    - The outreach has been good.

    - Does the ICTR provide the necessary tools to extract needed social determinant measures?
        - Yes
    - If not, what other tools would you like? Yes, and we are developing the tools. The tools are being modeled on what is available across the system.
- Institutional approval:
    - Do you think IRBs and PIs view social determinants differently, and if so, how?
        - Data trust is the bigger challenge. Sharing with the NNDC database is a bigger issue.
    - Have you seen problems in getting the collection of social determinant measures approved? If so, what kinds of problems? What happened?
- Do you have any other thoughts about these issues?
- New items to consider
        - IRB and Data trust are bigger issues.
        - Pulling information from another platform is a bigger issue.

■ APPENDIX B – DATA MATRIX AND COMMON VARIABLES

Figure App B1 – Data matrix that will be applied against common EPIC's social/behavioral data

What *Variable of interest*

- Variable Name: Click here to enter text.
- Variable Synonyms: Click here to enter text.
- UMLS ID #: Click here to enter text.
- Variable Type: ☐Genomic ☐ Clinical ☐ Behavioral / Psychological ☐ Social ☐ Environment

Whose *Variable exists for this patient denominator*

- Typically collected for these patients: Click here to enter text.
- Completeness (non-missing) rate (%):
  - Inpatient – all time Click here to enter text.
  - Inpatient – after 6/2016 Click here to enter text.
  - Outpatient – all time Click here to enter text.
  - Outpatient – after 6/2016 Click here to enter text.
- Notes about completeness: Click here to enter text.

When *Temporal aspects of the variable*

- First started to collect the variable: Facility: Click here to enter text. Date: Click here to enter text.
- Last started to collect the variable: Facility: Click here to enter text. Date: Click here to enter text.
- Date stopped collecting the variable: Date: Click here to enter text.
- Other significant dates/events: Click here to enter text.
- Frequency of variable collection: Click here to enter text.

Where *Location that variable is often collected*

- JHMI Location: Click here to enter text.
- Non-JHMI Healthcare Provider Location: Click here to enter text.
- Other Geographical Location: Click here to enter text.

Who *Person collecting the variables*

- Person usually collecting the variable: ☐Clinician ☐Admin Staff ☐Technician ☐Paramedic ☐Patient/Family
- Other person: Click here to enter text.

Data Management

- Data Provenance:
  - Epic Database: ☐Transactional (Chronicles) ☐Data Warehouse (Clarity) ☐Population Manage. (Cogito) ☐Other: Click here to enter text.
  - Epic Data Source: ☐Demographics ☐Encounters ☐Vital Signs ☐Other Flowsheet ☐Diagnosis ☐Problem List ☐Medication Order ☐Medication Reconciliation ☐Questionnaire ☐Specific Collection Form ☐Laboratory ☐Radiology/Imaging ☐Pathology ☐Clinical Notes ☐Other: Click here to enter text.
  - Other JHMI Source: Click here to enter text.
  - External Data Source: Click here to enter text.
- Data Type:
  - Data Structure: ☐Coded ☐Smart Data ☐Free Text ☐Other: Click here to enter text.
  - Coding Standard: ☐ICD ☐SNOMED ☐CPT ☐LOINC ☐HL7/FHIR ☐Other coding standard: Click here to enter text.
- Data Quality Comments:
  - Accuracy: Click here to enter text.
  - Completeness: Click here to enter text. (also see Whose)
  - Timeliness: Click here to enter text. (also see When)

Figure App B2 – Race Data Matrix

What *Variable of interest*

- Variable Name: Race
- Variable Synonyms: Click here to enter text.
- UMLS ID #: Click here to enter text.
- Variable Type: ☐Genomic ☐ Clinical ☐ Behavioral / Psychological ☒ Social ☐ Environment

Whose *Variable exists for this patient denominator*

- Typically collected for these patients: All patients upon admission
- Completeness (non-missing) rate (%):
  - All time Out of 5.1 million unique patients existing in Epic, 4.5 million indicated at least one race.
- Notes about completeness: See comments

When *Temporal aspects of the variable*

- First started to collect the variable: Facility: Click here to enter text. Date: Click here to enter text.
- Last started to collect the variable: Facility: Click here to enter text. Date: Click here to enter text.
- Date stopped collecting the variable: Date: N/A
- Other significant dates/events: Race data was backfilled from legacy systems when patient data was backfilled – April 2013
- Frequency of variable collection: Upon admission (inpatient and outpatient)

Where *Location that variable is often collected*

- JHMI Location: Click here to enter text.
- Non-JHMI Healthcare Provider Location: What does this mean? All Epic locations are considered JHMI locations, even JHCP.
- Other Geographical Location: Click here to enter text.

Who *Person collecting the variables*

- Person usually collecting the variable: ☐Clinician ☒Admin Staff ☐Technician ☐Paramedic ☐Patient/Family
- Other person: Click here to enter text.

Data Management

- Data Provenance:
  - Epic Database: ☐Transactional (Chronicles) ☐Data Warehouse (EDW) ☐Population Manage. (Cogito) ☐Other: Click here to enter text.
  - Epic Data Source: ☒Demographics ☐Encounters ☐Vital Signs ☐Other Flowsheet ☐Diagnosis ☐Problem List ☐Medication Order ☐Medication Reconciliation ☐Questionnaire ☐Specific Collection Form ☐Laboratory ☐Radiology/Imaging ☐Pathology ☐Clinical Notes ☐Other: Click here to enter text.
  - Other JHMI Source: Click here to enter text.
  - External Data Source: Click here to enter text.
- Data Type:
  - Data Structure: ☐Coded ☐Smart Data ☐Free Text ☒Other: Multi-select options: White or CaucasianBlack or African AmericanAmerican Indian or Alaska NativeAsianNative Hawaiian or Other Pacific IslanderOtherPatient RefusedUnknownTwo or More RacesDeclined to AnswerHispanic
  - Coding Standard: ☐ICD ☐SNOMED ☐CPT ☐LOINC ☐HL7/FHIR ☐Other coding standard: Click here to enter text.
- Data Quality Comments:
  - Accuracy: Click here to enter text.
  - Completeness: Click here to enter text. (also see Whose)
  - Timeliness: Click here to enter text. (also see When)

Figure App B3 – Ethnicity Data Matrix

What *Variable of interest*

- Variable Name: Ethnicity
- Variable Synonyms: Click here to enter text.
- UMLS ID #: Click here to enter text.
- Variable Type: ☐Genomic ☐ Clinical ☐ Behavioral / Psychological ☒ Social ☐ Environment

Whose *Variable exists for this patient denominator*

- Typically collected for these patients: All patients upon admission
- Completeness (non-missing) rate (%):
  - All time Out of 5.1 million unique patients existing in Epic, 2.5 million indicated an ethnicity.
- Notes about completeness: See comments

When *Temporal aspects of the variable*

- First started to collect the variable: Facility: Click here to enter text. Date: Click here to enter text.
- Last started to collect the variable: Facility: Click here to enter text. Date: Click here to enter text.
- Date stopped collecting the variable: Date: N/A
- Other significant dates/events: Ethnicity was backfilled from legacy systems when patient data was backfilled – April 2013
- Frequency of variable collection: Upon admission (inpatient and outpatient)

Where *Location that variable is often collected*

- JHMI Location: Click here to enter text.
- Non-JHMI Healthcare Provider Location: What does this mean? All Epic locations are considered JHMI locations, even JHCP.
- Other Geographical Location: Click here to enter text.

Who *Person collecting the variables*

- Person usually collecting the variable: ☐Clinician ☒Admin Staff ☐Technician ☐Paramedic ☐Patient/Family
- Other person: Click here to enter text.

Data Management

- Data Provenance:
  - Epic Database: ☐Transactional (Chronicles) ☐Data Warehouse (EDW) ☐Population Manage. (Cogito) ☐Other: Click here to enter text.
  - Epic Data Source: ☒Demographics ☐Encounters ☐Vital Signs ☐Other Flowsheet ☐Diagnosis ☐Problem List ☐Medication Order ☐Medication Reconciliation ☐Questionnaire ☐Specific Collection Form ☐Laboratory ☐Radiology/Imaging ☐Pathology ☐Clinical Notes ☐Other: Click here to enter text.
  - Other JHMI Source: Click here to enter text.
  - External Data Source: Click here to enter text.
- Data Type:
  - Data Structure: ☐Coded ☐Smart Data ☐Free Text ☒Other: Option menu: Hispanic or Latino; Not Hispanic or Latino; Patient Refused; Unknown
  - Coding Standard: ☐ICD ☐SNOMED ☐CPT ☐LOINC ☐HL7/FHIR ☐Other coding standard: Click here to enter text.
- Data Quality Comments:
  - Accuracy: Click here to enter text.
  - Completeness: Click here to enter text. (also see Whose)
  - Timeliness: Click here to enter text. (also see When)

Figure App B4 – Alcohol Usage Data Matrix

What *Variable of interest*

- Variable Name: Alcohol Use
- Variable Synonyms: Click here to enter text.
- UMLS ID #: Click here to enter text.
- Variable Type: ☐Genomic ☐ Clinical ☐ Behavioral / Psychological ☒ Social ☐ Environment

Whose *Variable exists for this patient denominator*

- Typically collected for these patients: All patients upon admission
- Completeness (non-missing) rate (%):
  - All time Out of 5.1 million unique patients existing in Epic, about 148,000 patients have reported having at least 1 alcoholic drink.
- Notes about completeness:

When *Temporal aspects of the variable*

- First started to collect the variable: Facility: Date: April 2013
- Last started to collect the variable: Facility: Click here to enter text. Date: Click here to enter text.
- Date stopped collecting the variable: Date: N/A
- Other significant dates/events:
- Frequency of variable collection:

Where *Location that variable is often collected*

- JHMI Location: Click here to enter text.
- Non-JHMI Healthcare Provider Location: What does this mean? All Epic locations are considered JHMI locations, even JHCP.
- Other Geographical Location: Click here to enter text.

Who *Person collecting the variables*

- Person usually collecting the variable: ☒Clinician ☐Admin Staff ☐Technician ☐Paramedic ☐Patient/Family
- Other person:

Data Management

- Data Provenance:
  - Epic Database: ☒Transactional (Chronicles) ☐Data Warehouse (EDW) ☐Population Manage. (Cogito) ☒Other: Social History
  - Epic Data Source: ☐Demographics ☐Encounters ☐Vital Signs ☐Other Flowsheet ☐Diagnosis ☐Problem List ☐Medication Order ☐Medication Reconciliation ☐Questionnaire ☐Specific Collection Form ☐Laboratory ☐Radiology/Imaging ☐Pathology ☐Clinical Notes ☐Other: Click here to enter text.
  - Other JHMI Source: Click here to enter text.
  - External Data Source: Click here to enter text.
- Data Type:
  - Data Structure: ☐Coded ☐Smart Data ☐Free Text ☒Other: Number of alcoholic drinks a week – numeric field
  - Coding Standard: ☐ICD ☐SNOMED ☐CPT ☐LOINC ☐HL7/FHIR ☐Other coding standard: Click here to enter text.
- Data Quality Comments:
  - Accuracy: Click here to enter text.
  - Completeness: Click here to enter text. (also see Whose)
  - Timeliness: Click here to enter text. (also see When)

Figure App B5 – Depression Data Matrix

What *Variable of interest*

- Variable Name: Depression Screening
- Variable Synonyms: PHQ-2, PHQ-9
- UMLS ID #: Click here to enter text.
- Variable Type: ☐Genomic ☐ Clinical ☒ Behavioral / Psychological ☐ Social ☐ Environment

Whose *Variable exists for this patient denominator*

- Typically collected for these patients: All patients upon admission
- Completeness (non-missing) rate (%):
  - All time Out of 5.1 million unique patients existing in Epic, about 300,000 patients have been screened for depression at least once.
- Notes about completeness:

When *Temporal aspects of the variable*

- First started to collect the variable: Facility: JHCP and other general practice clinics Date: April 2013
- Last started to collect the variable: Facility: Click here to enter text. Date: Click here to enter text.
- Date stopped collecting the variable: Date: N/A
- Other significant dates/events:
- Frequency of variable collection: Upon admission (inpatient and outpatient)

Where *Location that variable is often collected*

- JHMI Location: Click here to enter text.
- Non-JHMI Healthcare Provider Location: What does this mean? All Epic locations are considered JHMI locations, even JHCP.
- Other Geographical Location: Click here to enter text.

Who *Person collecting the variables*

- Person usually collecting the variable: ☒Clinician ☐Admin Staff ☐Technician ☐Paramedic ☐Patient/Family
- Other person:

Data Management

- Data Provenance:
  - Epic Database: ☒Transactional (Chronicles) ☐Data Warehouse (EDW) ☐Population Manage. (Cogito) ☐Other: Click here to enter text.
  - Epic Data Source: ☐Demographics ☒Encounters ☐Vital Signs ☐Other Flowsheet ☐Diagnosis ☐Problem List ☐Medication Order ☐Medication Reconciliation ☒Questionnaire ☐Specific Collection Form ☐Laboratory ☐Radiology/Imaging ☐Pathology ☐Clinical Notes ☐Other: Click here to enter text.
  - Other JHMI Source: Click here to enter text.
  - External Data Source: Click here to enter text.
- Data Type:
  - Data Structure: ☐Coded ☐Smart Data ☐Free Text ☒Other: questionnaires
  - Coding Standard: ☐ICD ☐SNOMED ☐CPT ☐LOINC ☐HL7/FHIR ☐Other coding standard: Click here to enter text.
- Data Quality Comments:
  - Accuracy: Click here to enter text.
  - Completeness: Click here to enter text. (also see Whose)
  - Timeliness: Click here to enter text. (also see When)

Figure App B6 – Tobacco Use Data Matrix

What *Variable of interest*

- Variable Name: Tobacco Use
- Variable Synonyms: Click here to enter text.
- UMLS ID #: Click here to enter text.
- Variable Type: ☐Genomic ☐ Clinical ☐ Behavioral / Psychological ☒ Social ☐ Environment

Whose *Variable exists for this patient denominator*

- Typically collected for these patients: All patients upon admission
- Completeness (non-missing) rate (%):
  - All time: Out of 5.1 million unique patients existing in Epic, 1.4 million indicated provided data on tobacco use.
- Notes about completeness: See comments

When *Temporal aspects of the variable*

- First started to collect the variable: Facility: Click here to enter text. Date: Click here to enter text.
- Last started to collect the variable: Facility: Click here to enter text. Date: Click here to enter text.
- Date stopped collecting the variable: Date: N/A
- Other significant dates/events:
- Frequency of variable collection: Upon admission (inpatient and outpatient)

Where *Location that variable is often collected*

- JHMI Location: Click here to enter text.
- Non-JHMI Healthcare Provider Location: What does this mean? All Epic locations are considered JHMI locations, even JHCP.
- Other Geographical Location: Click here to enter text.

Who *Person collecting the variables*

- Person usually collecting the variable: ☒Clinician ☒Admin Staff ☐Technician ☐Paramedic ☐Patient/Family
- Other person:

Data Management

- Data Provenance:
  - Epic Database: ☒Transactional (Chronicles) ☐Data Warehouse (EDW) ☐Population Manage. (Cogito) ☐Other: Click here to enter text.
  - Epic Data Source: ☐Demographics ☐Encounters ☐Vital Signs ☐Other Flowsheet ☐Diagnosis ☐Problem List ☐Medication Order ☐Medication Reconciliation ☐Questionnaire ☐Specific Collection Form ☐Laboratory ☐Radiology/Imaging ☐Pathology X Social History ☐Clinical Notes ☐Other: Click here to enter text.
  - Other JHMI Source: Click here to enter text.
  - External Data Source: Click here to enter text.
- Data Type:
  - Data Structure: ☐Coded ☐Smart Data ☐Free Text ☒Other: Option menu: Current Every Day Smoker, Current Some Day Smoker, Former Smoker, Heavy Tobacco Smoker, Light Tobacco Smoker, Never Assessed, Never Smoker, Passive Smoke Exposure – Never Smoker; Smoker, Current Status Unknown; Unknown if Ever Smoked.
  - Coding Standard: ☐ICD ☐SNOMED ☐CPT ☐LOINC ☐HL7/FHIR ☐Other coding standard: Click here to enter text.
- Data Quality Comments:
  - Accuracy: Click here to enter text.

Figure App B7 – Residence Zip Code Data Matrix

What *Variable of interest*

- Variable Name: Patient Zip Code
- Variable Synonyms: Click here to enter text.
- UMLS ID #: Click here to enter text.
- Variable Type: ☐Genomic ☐ Clinical ☐ Behavioral / Psychological ☒ Social ☐ Environment

Whose *Variable exists for this patient denominator*

- Typically collected for these patients: All patients upon admission
- Completeness (non-missing) rate (%):
  - All time Out of 5.1 million unique patients existing in Epic, only about 300,000 patients do not have a zip code indicated in Epic.
- Notes about completeness: See comments

When *Temporal aspects of the variable*

- First started to collect the variable: Facility: Click here to enter text. Date: Click here to enter text.
- Last started to collect the variable: Facility: Click here to enter text. Date: Click here to enter text.
- Date stopped collecting the variable: Date: N/A
- Other significant dates/events:
- Frequency of variable collection: Upon admission (inpatient and outpatient)

Where *Location that variable is often collected*

- JHMI Location: Click here to enter text.
- Non-JHMI Healthcare Provider Location: What does this mean? All Epic locations are considered JHMI locations, even JHCP.
- Other Geographical Location: Click here to enter text.

Who *Person collecting the variables*

- Person usually collecting the variable: ☐Clinician ☒Admin Staff ☐Technician ☐Paramedic ☐Patient/Family
- Other person:

Data Management

- Data Provenance:
  - Epic Database: ☒Transactional (Chronicles) ☐Data Warehouse (EDW) ☐Population Manage. (Cogito) ☐Other: Click here to enter text.
  - Epic Data Source: ☒Demographics ☐Encounters ☐Vital Signs ☐Other Flowsheet ☐Diagnosis ☐Problem List ☐Medication Order ☐Medication Reconciliation ☐Questionnaire ☐Specific Collection Form ☐Laboratory ☐Radiology/Imaging ☐Pathology ☐Clinical Notes ☐Other: Click here to enter text.
  - Other JHMI Source: Click here to enter text.
  - External Data Source: Click here to enter text.
- Data Type:
  - Data Structure: ☐Coded ☐Smart Data ☒Free Text ☐Other:
  - Coding Standard: ☐ICD ☐SNOMED ☐CPT ☐LOINC ☐HL7/FHIR ☐Other coding standard: Click here to enter text.
- Data Quality Comments:
  - Accuracy: Click here to enter text.
  - Completeness: Click here to enter text. (also see Whose)
  - Timeliness: Click here to enter text. (also see When)

■ APPENDIX C – EXTRACTING DATA FROM EPIC

• *CCDA Data Request Guidance*

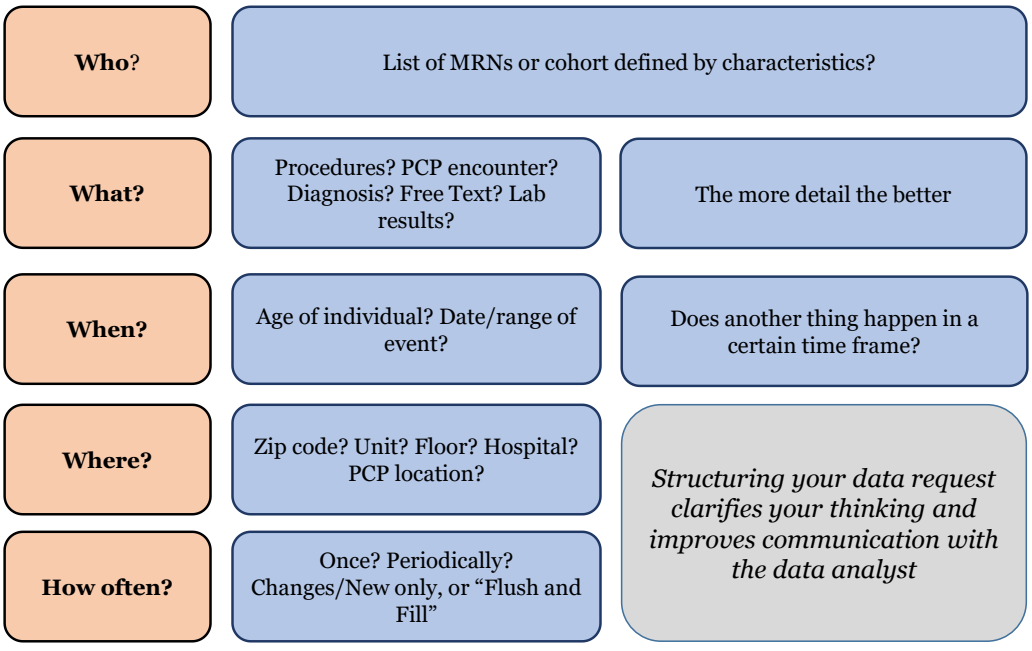
About this Document

Thank you for using the Center for Clinical Data Analysis (CCDA) to meet your data research needs. This document was prepared to explain details and caveats regarding the data delivered to you. If you have further questions about your data, please contact Bonnie Woods (Bonnie.Woods@jhu.edu) for follow-up.

About EPIC Data

We recommend that you closely work with CCDA in translating your high concept research data questions and asks into actionable data collection queries. This will require identifying the "who, what, when, where, and how often" attributes of the data that can answer your specific research question (Figure App B1)

Figure App C1 – Structuring a data request



Note that histoical data for all patients do not exists in EPIC. While transitioning to the EPIC EMR, some of the historical data were not imported (Figure App B2). Different facilities (e.g., hospitals) migrated to EPIC on different dates making data availability heteregenous across them (Figure App B3).

Figure App C2 – Historical data backloaded into EPIC

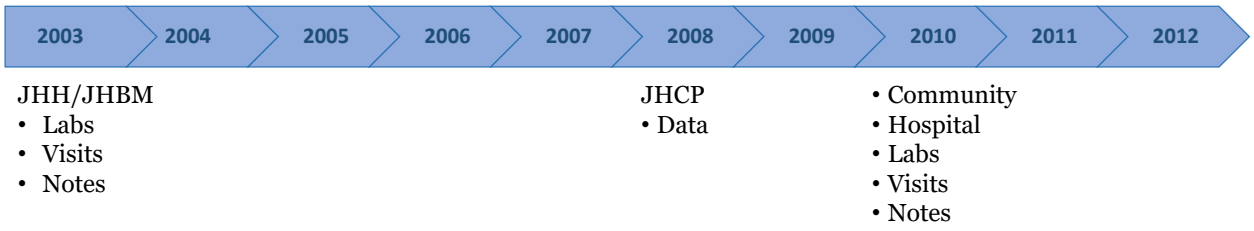
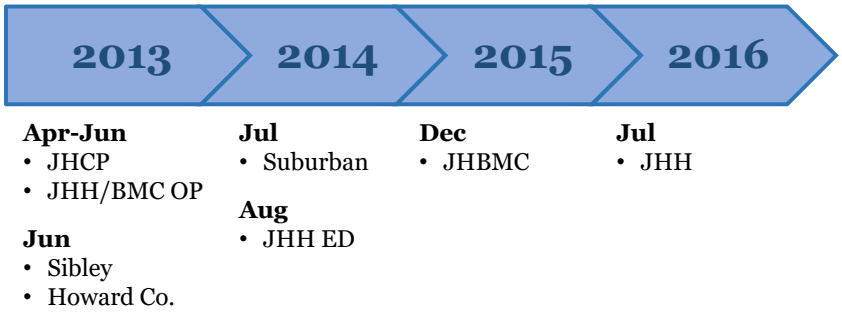


Figure App C3 – Rollout of EPIC in various settings/facilities



About Your Data

Delivered to a secure location: Your data has been placed on a file server which is approved for delivery of PHI (\\win.ad.jhu.edu\cloud\yourprojectfolder[TBD]$).

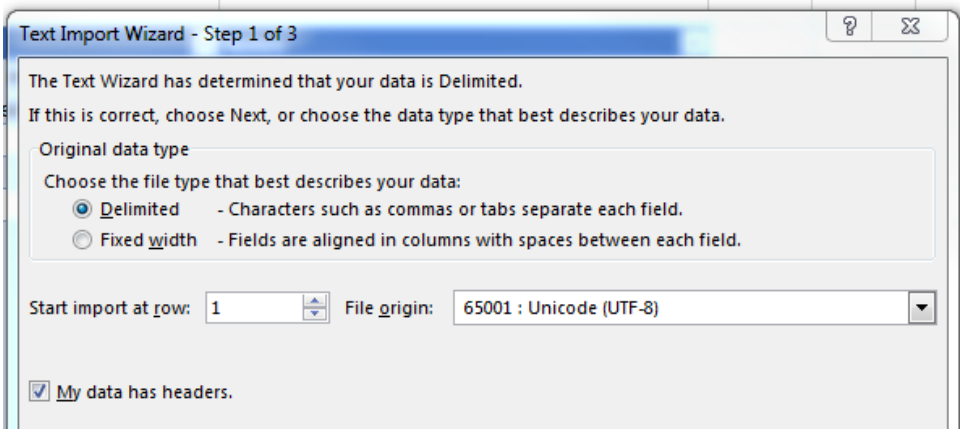
To meet your responsibility for the security of this data, you should consider this location for your work. If space constraints or other concerns cause you to considering moving this data to do your analysis, you are responsible for doing so in compliance with the Data Use Agreement (DUA) you signed, and policies of Johns Hopkins Medicine. CCDA is available to help you evaluate your needs and put you in touch with enterprise resources to ensure the security of your research data.

File Format

Your data was exported in pipe-delimited format (.txt) instead of Excel (.xslx) due to the limitations of Excel with large data sets. To open the files in Excel, follow the steps below:

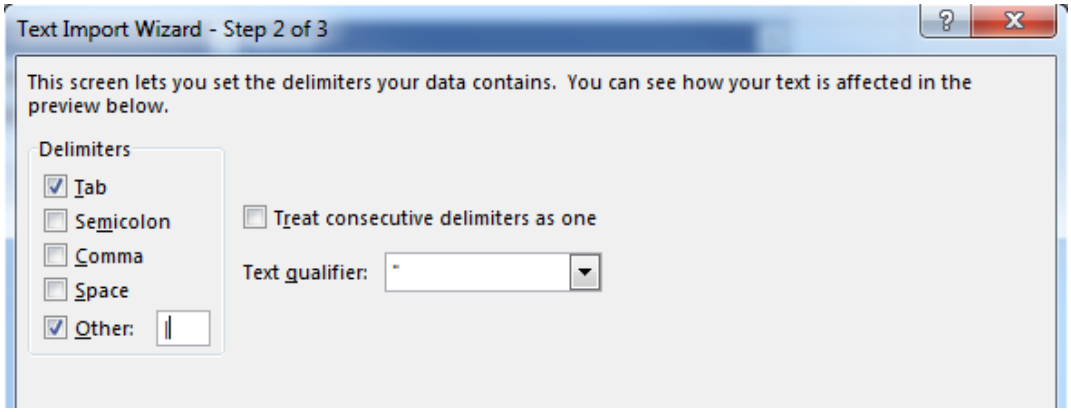
1. Select Delimited from the original file type, and select the “My data has headers” option button. Click Next to continue.

Figure App C4 – Importing CCDA data into Excel (Part 1)



2. Select the "Tab" and "Other" option buttons, and type the pipe (|) in the text area next to "Other". (Pipe is the shift character above the Enter key.) Click Next to continue.

Figure App C5 – Importing CCDA data into Excel (Part 2)



3. You can preview your data by clicking the Finish button.

Patient Inclusion and Exclusion Criteria

Inclusion:

- Adult patients (>= 21 years of age at the time of the extraction)
- For first extraction: Having a primary care clinic office visit within the last six months (at date of extraction) at JHCP Frederick
- Having an ethnicity of Hispanic **or** a race of either White or African American (*Note: if the patient selected White and African American, we returned one or the other, not both.)*
- Having either a visit diagnosis or a problem list diagnosis of HTN (ICD 9 – 401.X; ICD 10 – I10.X)
- Having a Systolic BP ≥ 140 mmHg or diastolic BP ≥ 90 mmHg on the last BP recorded at the most recent encounter (at JHCP Frederick)

- Having at least one of the following ICD codes on the problem list or a visit diagnosis:
    - ICD-9: 402.XX, 410.XX-414.XX, 429.2XX, 305.1XX, 250.XX, 272.XX or 296.2XX, 296.3XX, 311.XX
    - ICD10: I25.XX, F17.XX, E10.XX, E11.XX, E78.XX, F32.XX or F33.XX

Exclusion:

- Patients known to be deceased. If a patient dies at a non-JHM facility and the family does not make JHM aware of the death, EPIC will not indicate that the patient is deceased.
- Patients who have an ICD-9 code of 585.6 or an ICD-10 code of N18.6 (end stage renal disease) on the problem list or visit encounter. These ICD codes do not need wildcards (X) after the code because there are no subcategories for these codes.

Patient Demographics: Primary Care Provider

This data element is not always collected or modified accurately. We provided the PCP, NPI, and PCP Department that was entered into EPIC at the time of the data extract.

Patient Encounters

All patient encounters are JHCP Frederick office visits with encounter dates within 12 months of the data extract run date.

The payor information delivered in the encounters file is the patient's primary insurance recorded at the time of the encounter.

There is no Plan Effective Date recorded in the Clarity reporting database at this time. We will contact our EPIC team to ask them to investigate this issue.

The Blood Pressure readings are the last BP vitals recorded at the encounter.

Lab values included

Most recent random glucose, fasting glucose, hemoglobin A1c, LDL, HDL, total cholesterol, triglycerides, eGFR. The study team was sent a full list of base names and common names of these labs to exclude or include. If the study team wants to add or remove values, the CCDA will make the change and re-run the lab extract.

Depression Screening

The extract file for depression screening contains the PHQ-9 questions and answers for each encounter occurring within the last 12 months of the data extract run date. The PHQ-9 questionnaire uses the AMB PHQ-9 DEPRESSION SCALE template.

Social and Behavioral Data

[To Be Completed]

- ***CCDA Extract Specification***

CCDA will need the specific information about the patient cohort/denominator of interest, source of data, and other adnimistrative information before a query can be executed to extract data (including social and behavioral data). Table B1 lists some of the information that CCDA will collect and put together before a data pull can be executed.

Table App B1 - Extract background and status

JIRA	[CCDA-xxx]
Study PI	
Study Title	
Contact	[if different from PI]
Date	
Extract purpose	[brief description of study as well as purpose for extracting data]
Current IRB status	[e.g., IRB number, IRB name (IRB-X, etc.), and status (approved, pending)
Funding available	[enter cost center number if available]
Extract frequency	[one-time, weekly, monthly, etc.]
Data Source	[EPIC, SCM, CaseMix, EPR2020, etc.]
Extract Structure	[Excel, pipe-delimited, CSV, SQL tables – we are starting to send everything as pipe-delimited to avoid errors with large data sets and Excel]
Data Delivered To	[server name, share name – or JHBox, Enterprise NAS, etc.]
Data Shared with external entity?	[Include information on researcher's intent to share outside of JHM. This includes corporate sponsors and multi-site studies. Also include information on what data elements are proposed to be shared and in what format (PHI, limited data set, etc.)]
Work Estimate	[estimate in hours]

Inclusion criteria - Only patients with the following criteria will be included in the extract results: [to be filled]

Exclusion criteria - Patients with the following criteria will be excluded from the extract results: [to be filled]

Extract sections and format: The extract output will consist of x section(s): Add sections (table) to represent one-to-many or many-to-many relationships.

Table App B2 – Data element relationships

Data Element	Notes
[element 1]	[notes]
[element 2]	[notes]
[element 3]	[notes]

Comments:

1. The CCDA will conduct a review of the IRB protocol to ensure that requested data match what was approved by the IRB.
2. "Data Use Agreement" (DUA) needs to be signed by PI before we can begin work.
3. This project may need to be reviewed by the Data Trust Research Sub-council, depending on cohort size.
4. Mr. Darren Lacey (dll@jhu.edu), Johns Hopkin's Chief Information Security Officer, needs to confirm the security of the destination server before data can be delivered to any server.
5. Data requests for Johns Hopkins Community Physician (JHCP) patient data will need to be approved by the JHCP data committee. Contact Jennifer Bailey (jbailey@jhmi.edu) for more information.

- ***Data Trust Review of Research Data Requests FAQ***

  - *What is the JHM Data Trust Council?*

The Data Trust Council (DTC) governs JHM data (data in JHM clinical, health plan, and business systems), making such data readily available for appropriate use while protecting patient privacy and maintaining data security. The DTC has subcouncils, each with a different responsibility (*e.g.*, research use, quality improvement, security), to review and approve data requests and propose policies. The actions and oversight of the DTC were authorized in 2016 when the participating JHM provider entities (including JHH, Suburban Hospital, Sibley Memorial Hospital, Howard County General Hospital, and JHCP) and health plans signed the JHM Data Trust Policy, establishing the DTC and giving it authority to oversee JHM data use and approve data requests.

Note that all Hopkins data, even if not subject to Data Trust oversight (*e.g.*, data collected solely for research, not used for patient care, and not stored in any clinical system), must still be stored, used, and disclosed in compliance with the appropriate agreements regarding data use as well as IRB and Johns Hopkins IT policies and requirements, which include encryption, server security, and access controls.

The "Data Trust Research Data Subcouncil" develops policy and reviews requests for *research* uses of JHM data. Hopkins IT and security experts, working with the "Center for Clinical Data and Analytics" (CCDA), help the Data Trust Research Data Subcouncil assess technical security, access controls, and Deidentification protocols for specific projects.

  - *Do all research requests for JHM Data require review?*

No. Many smaller projects require no review. Ordinarily, if a retrospective chart review involves less than 500 records and the IRB application contains an acceptable data security plan, upon IRB approval the researcher may seek data from the CCDA without Data Trust review. The CCDA may review the researcher's deidentification protocol (if applicable) to confirm that it meets HIPAA standards. If a project involves PHI, limited data sets, or "sensitive" deidentified data (e.g., genomic data, volumetric neuroimages), data sharing with collaborators or institutions outside the JHM covered entity it requires a written agreement with appropriate data use terms. Consult the "Office of Research Administration" (ORA) to determine whether data use terms are already included in any contract or funding agreement for the study.

Note about clinical trials: For most sponsored clinical trials, the subjects give written consent/HIPAA authorization and ORA has negotiated a contract with data use terms. These studies generally do not require Data Trust review.

Note about vendors: All vendors providing services for a study (e.g., cloud storage, data abstraction or analysis) must be in a contractual relationship with JH. Vendor contracts must receive legal review prior to signature.

- *Which research projects require Data Trust review?*

Projects meeting any of the following criteria require Data Trust review.

- Involving data is going to a commercial third party (excluding sponsored research agreements of less than 500 records).
- Involving sponsored chart reviews involving more than 499 records and a waiver of consent.
- Involving 500 or more records that will be shared with a third party or transferred outside the JHM firewall.
- Involving a live data feed from an enterprise clinical system.
- Involving data collection via an app.
- Referred by the IRB, CCDA, or other data stewards due to concerns about size, sensitivity or security.

Contact Valerie Smothers (vsmothers@jhmi.edu) to request Data Trust review.

- *Do I need IRB approval before contacting the Data Trust?*

No. It is possible to be working with the IRB and Data Trust Research Data Subcouncil at the same time, but approval of the final version by both parties is necessary. Unless the project has clear data definitions and a strong data security plan, the PI may save time by consulting the Data Trust Research Data Subcouncil or CCDA when designing the protocol.

- *May I transfer data without an agreement? (IRB approved or deidentified)*

No. With limited exceptions for research consultations, all transfers of data (including deidentified data) outside Hopkins for research use must occur under an appropriate legal agreement, and all vendors must be in a contractual relationship with Hopkins.

Before approving a data request the Data Trust Research Data Subcouncil may refer the PI to ORA or Hopkins attorneys to confirm that the necessary agreements are in place.

- *What do I need to know about deidentification?*

Complete deidentification of data is rarely achievable. Often, investigators fail to realize that dates, zip codes, or similar fields are considered identifiers.

Partial deidentification can be accomplished using techniques such as date shifting and hashing of identifiers. Even if all data elements that are defined by HIPAA as Protected Health Information are removed or obscured there is sometimes a risk that data could be reidentified if joined with an external data set. Therefore, deidentification is often not sufficient to protect privacy and data security. Deidentification of Hopkins data should be done at Johns Hopkins whenever possible, and the Data Trust offers expert help and support for deidentification. If deidentification is happening elsewhere, investigators will be asked by the Data Trust to provide

the deidentification protocol and identify any third parties (e.g., data managers, cloud vendors, app developers) who will receive data.

- ***Structure of Data Trust and Analytic Teams***

The Data Trust Council is responsible for overall governance of patient and health plan member-related data stored in the clinical enterprise systems of Johns Hopkins Medicine entities, including development of policies to ensure the quality, accessibility and use of data for appropriate purposes. The policies being put in place will ensure the quality and accessibility of that data. The council will also oversee the process for those requesting data for research or operations. The council has several sub-councils that help it achieve its goals. See Data Trust Organization for details (Figure App B6).

Figure App C6 – Data Trust teams

The Operations Team is a central team that will support the development of shared Data Trust infrastructure and coordinated analytics. It will play a coordinating role across the 10 approved Analytic Teams.

Analytic Teams work to coordinate analytic efforts across Johns Hopkins Medicine within a defined scope. They help reduce redundant efforts and encourage use of common infrastructure. Analytic Teams also play a role in building data flows to efficiently support analytic needs. These teams will consider and fulfill quality, operational and research-related requests for data. The teams focus on:

- Ambulatory operations

- Ambulatory quality
- Hospital quality
- Hospital operations
- Hospital utilization management
- Finance-integrated analytics
- Population health
- Research/Center for Clinical Data Analysis (CCDA)
- Technology Innovation Center
- Planning and market analysis

Follow these links to access additional information about the Data Trust and see guidelines for requesting access and data.

- Operations and guiding principles
- Data Trust policies
- Requesting access to the Data Trust infrastructure
- Requesting data from an Analytic Team

Analytic Teams approve access to components of the Data Trust Infrastructure for analysts working within their purview. They also consider and fulfill quality, operational and research-related requests for data. Many Analytic Teams operate virtually and may report to different individuals. Below is a list of the Analytic Teams:

- http://intranet.insidehopkinsmedicine.org/data_trust/analytic_teams/

- ***EPIC's Slicer Dicer FAQ***

Slicer Dicer is a built-in EPIC tool to explore patient denominators based on criteria specified by users interactively (i.e., counting patients fitting a condition). Following are some of the common questions and answers about Slicer Dicer received and answered by CCDA staff members in the past.

Q: What patients are included in SlicerDicer?

A: Any JHMI patient that has had an appointment or an admission in EPIC since April 2013 will be included.

Q: How does SlicerDicer treat "null" values versus "unknown" values?

A: If a data element was not entered for a patient (e.g., race, ethnic group, marital status), the data element's value is missing (i.e., NULL value), and SlicerDicer will not include these values in a count. If a data element is marked as unknown (i.e., unknown race, unknown gender), these values will be included in a count and can be split as "unknown".

Q: How can I select multiple values for my criterion?

A: When you add criteria to your search ("Add Criteria" button, then select the criterion), you will need to select each value one at a time by typing the value in the white search box. There is currently no functionality for selecting multiple values at the same time for a criterion (e.g., multiple diagnoses).

Q: Why doesn't the result change when I change the timeframe for some data types?

A: The SlicerDicer timeframe will not affect the initial "All Patients" group. At least one criterion must be added before the timeframe will have an effect on the query. Any data type marked as "Current" will not be affected by the timeframe.

Q: Why can't I filter labs by lab value?

A: This is due to the variances in reference ranges across our 9 different lab data sources. If you need to filter labs by a specific lab value, then visit https://cscop.jhmi.edu/jira/browse/DT to request help from an analytics team.

Q: Why don't I see counts less than 10 when I search "All Patients"?

A: To protect patient privacy, for searches resulting in fewer than 10 patients you will not get an exact number. You will only see that there are 10 or fewer records. When searching "My Patients", you will see the exact counts for your search results.

Q: Why aren't ophthalmic surgeries available as a criterion under "Procedures"?

A: Surgeries performed in Ophthalmology Ambulatory Surgery Centers are not currently ordered in EPIC and are therefore not available as search criteria in the Procedures folder. If an order is placed within EPIC for a procedure, then it can be queried by adding a condition for the procedure.

Q: Why can't I filter my criteria by country?

A: Country is not an available filter at this time. This ability has been added to the enhancements list to be available in a future version of SlicerDicer.

Q: How can I apply my filter to multiple populations at the same time?

A. There is currently no way to add the same filter to multiple populations simultaneously. The approach to take is to set up your population and then split on a particular filter.

Q: Why didn't Slicer Dicer refresh the results after I added criteria to my split populations?

A: Once populations are split, new or deleted criteria must be manually applied to each split population. This cannot be done simultaneously.

Q: When entering criteria, the "More matches exist" message appears, but I don't see a way to load more results. Why?

A: When searching for values to add to your criteria, type in at least three characters to view the "Load More" link. This makes it easier for the system to narrow down your results.

Q: How many different populations can I view in Slicer Dicer at one time?

A: A total of 10 different populations are able to be viewed in Slicer Dicer simultaneously. More than 10 populations make it difficult to compare populations, especially on smaller monitors.

Q: What is the logic behind the "Pregnancy" criteria?

A: The Pregnancy criteria are based on episodes in EPIC. An episode allows providers to collect and view information from several related encounters through flowsheets and/or special reports in the Episodes activity. In 2013, when EPIC first went live, use of the pregnancy episode was very low. Pregnancy episode frequency improved to a little over 50% of the population in 2014. In 2015, it rose to approximately 85%. Currently the use of the pregnancy episode workflow is on par with pregnancy diagnoses.

Q: How often is Slicer Dicer data refreshed from EPIC's production system?

A: Slicer Dicer is refreshed nightly from the Production environment but represents the previous day's data. Use "Reporting Workbench" or "Clarity Reports" for real-time operational reporting needs.

Q: Why is Slicer Dicer taking so long to show my population?

A: If your query's date range is 3 or more years, it might take up to a minute or longer to find the results from the query. A typical query of 2 or less years should show results quickly. If that's not the case, please open an EPIC Help Desk (a.k.a., Remedy) ticket.

Q: How can I report issues I'm having with using Slicer Dicer?

A: Open an EPIC Help Desk (Remedy) ticket and an analyst will help trouble-shoot the issue.