1A) National-level Data Examples: Free or Inexpensive

- **NHANES - National Health and Nutrition Examination Survey (NHANES).** Selected diseases and conditions including those undiagnosed or undetected - Nutrition monitoring-Environmental exposures monitoring- Infectious disease monitoring-Overweight and diabetes--Respiratory diseases--Hypertension and cholesterol—Hearing. Income and poverty index—Education—Occupation---Type of living quarters---Social services---Birthplace---Acculturation questions re: language usually spoken at home. The NHANES I target population is the civilian non-institutionalized population ages one to 74 years of age residing in the continental United States. The NHANES II target population was the civilian non-institutionalized population 6 months-74 years of age residing in the United States, including Alaska and Hawaii. Race information for NHANES I and NHANES II was determined primarily by interviewer observation.

- **BRFS - The Behavioral Risk Factor Surveillance System (BRFSS)** is an ongoing data collection program designed to measure behavioral risk factors for adults 18 years of age or older living in households. The BRFSS is administered and supported by the Centers for Disease Control and Prevention's Behavioral Risk Factor Surveillance Branch and is a collaborative project of the Centers for Disease Control and Prevention (CDC) and U.S. states and territories. The BRFSS objective is to collect uniform, state-specific data on preventive health practices and risk behaviors that are linked to chronic diseases, injuries, and preventable infectious diseases that affect the adult population. Factors assessed by the BRFSS include tobacco use, health care coverage, HIV/AIDS knowledge and prevention, physical activity, and fruit-and-vegetable consumption (Centers for Disease Control and Prevention [CDC], 2012). The data are free and can be downloaded from the CDC website.

- **MCBS - The Medicare Current Beneficiary Survey (MCBS)** is a continuous multi-purpose, panel survey of the Medicare population sponsored by the Center for Medicare and Medicaid Services (CMS). The sample for the MCBS is drawn from the CMS' Medicare enrollment file with current data collected during Round 61, September through December of 2011. These data are augmented with Medicare claims and administrative data for calendar year 2011. A complex, multi-stage, stratified sampling design is used to obtain a nationally representative sample of all Medicare beneficiaries. The United States is divided geographically by counties into 107 primary sampling units and then further divided into 1,258 clusters by postal zip codes for the 2011 survey. The beneficiaries included in the 2011 Access to Care File consist of a random cross-section of all beneficiaries who were enrolled in one or both parts of the Medicare program as of January 1, 2011 and were alive and enrolled at the time of interview during the 2011 fall round, September through December (CMS, 2011). The MCBS can be purchased for $XXX, and/or investigators may be able to collaborate with others on campus who have use of these data.

1B) National Data Available at MUSC

A number of national data sets may be available to MUSC faculty and student researchers through The MUSC College of Health Professions’ (CHP) Archival Data Analysis Collaborative (ADAC). Dean Lisa Saladin and Dr. Zoller, Chair of the Department of Health Leadership and Management in CHP have combined resources and provided all CHP faculty and students with an opportunity to explore comparative effectiveness research questions and evidence-based practice patterns in large national...
data bases. Some of these data are also available to MUSC faculty and students in other colleges. These types of data are not usually available to individual faculty without major external funding, however, the CHP leadership decided to provide support for unfunded exploratory research studies for faculty and students by purchasing institutional access to these valuable real-life “Big Data”. These big data can support research to examine questions such as: 1) effect of reimbursement on use of rehabilitation interventions; 2) variations in practice patterns by provider for complex chronic conditions; or 3) 10 year trends in the use of inpatient care modalities. The data includes one year of Medicare 5% data sample (2.7 million patients), 10 years of HCUP hospital discharge data from 15 US states (150 million admissions), 10 years of a national sample of all US hospital admissions (72 Million admissions), and best of all 3 years of Market Scan data containing physician, hospital, nursing home and pharmacy records for 4 million commercially insured patients.

• **Market Scan® Data** The College of Health Professions Archival Data Analysis Collaborative (ADAC) has a one-year institutional license to 3 years of Market Scan® data for 2010-2012 containing physician, hospital, nursing home and pharmacy records for 4 million commercially insured patients. Any MUSC faculty member or student may use these data through collaboration with ADAC investigators. These data must be accessed before May 31, 2015.

• **HCUP - The Healthcare Cost and Utilization Project**, sponsored by the Agency for Healthcare Research and Quality (AHRQ), includes the State Inpatient Database (SID). This state-specific database contains clinical and nonclinical inpatient discharge information from all payers that can be used to analyze cost and utilization. These data are anonymous, with all relevant patients and provider identifiers removed. These limited data sets can be purchased (at $35 per year) and require only certification that the researcher has completed the online training process and signed the data agreement statement. For IRB purposes, this type of research is classified as NON-HUMAN RESEARCH, because no individual protected health information remains on the data. These data may be accessible to MUSC researchers through the College of Health Professions (CHP) Archival Data Analysis Collaborative (ADAC).

• **NIS - Data from the National Inpatient Sample** includes discharge data from more than 1,000 hospitals in 45 states, which include 96% of the United States population. The sample contains all payer hospital discharge data based on the Uniform Budget Act 2004 discharge form that is filed by all US acute care hospitals. The data are part of the AHRQ HCUP data family. Data from 2000-2012 may be available through the CHP ADAC.

• **Medicare 5% Sample for 2012.** A 5% sample of Medicare patient bills for year 2012 may be available for certain research questions through collaboration with CHP ADAC investigators. This database contains bills for hospital, outpatient, MD visits, durable medical equipment and therapy visits for Medicare patients. There are no prescription pharmacy data in this database.

2) **State and MUSC Medical Record Data**

• **EDW - MUSC Enterprise Data Warehouse** [https://sctr.musc.edu/index.php/cdw](https://sctr.musc.edu/index.php/cdw) contains most of the data in the electronic health record used for patients seen at MUSC. Currently an estimated 1.3 Million patient records are included in the database. You will find links to data dictionaries and tutorials and a link to the self-service access at the above website. The Biomedical Informatics Center and the
OCIO EDW team are working on upgrading the EDW underlying architecture and the self-service interface to a tool that is more researcher-friendly. The governance of the EDW includes oversight of both research and non-research activities, is being led by the Chief Research Informatics Officer (CRIO) and the hospital's Chief Analytics Officer (CAO). The Governance Committee has been charged by the Dean of the College of Medicine.

Data are made available to researchers through one of two IRB approved methods:
1) **Self-service Access**: via a click through data use assurance. Only de-identified, aggregate data is made available using this method. The objective of this method is to allow investigators to query aggregate clinical data in order to perform basic hypothesis testing using, cohort identification and analysis, and correlative studies.
2) **Brokered Access**: the objective of method is to allow investigators to request more detailed or complex queries, which may include Protected Health Information (PHI) from the EDW. Investigators have to have IRB approval if PHI is needed. The data is requested through SCTR via SPARC Request. Data requests are reviewed by a Data Request Committee to ensure regulatory and IRB protocol compliance. Once approved, the requests are fulfilled by the Honest Broker and the data is transmitted to the investigator through a secure communication.

- **HSSC CDW - HSSC Clinical Data Warehouse**
  
http://www.healthsciencessc.org/cdw.asp In 2013, Health Sciences South Carolina (HSSC), a statewide research collaborative of the largest health systems and research intensive universities in South Carolina, publicly unveiled a multi-institutional, integrated Clinical Data Warehouse (CDW). The HSSC CDW captures, harmonizes and aggregates existing clinical and administrative data from multiple, independent health systems: Greenville Health System (GHS), Palmetto Health (PH) and the Medical University of South Carolina (MUSC). Spartanburg Regional Healthcare System (SRHS) will follow soon. The governance of the CDW includes oversight of both research and non-research activities, is based on a Data Collaboration Agreement (DCA) signed by the HSSC Board members (CEOs of the participating institutions). The Governance Committee is composed from senior leadership at participating institutions. The governance includes various working groups that oversee the continued development of the CDW, determine data access policies, ensure regulatory compliance and provide risk assessments. Data from EMRs on medical encounters, diagnoses, procedures are available from 2007 to present. Data on inpatient medication orders, and inpatient medication administration are available from 2012 to present. Data from 3,461,293 distinct patients are housed in this database.

Data are made available to researchers through one of two IRB approved methods:
1) **i2b2 Access**: via a click through data use agreement, de-identified, aggregate data. The objective of this method is to allow investigators to query aggregate clinical data in order to perform basic hypothesis testing using, cohort identification and analysis, and correlative studies, or 2) **Brokered Access**: the objective of method is to allow investigators to request more detailed or complex queries, which may include Protected Health Information (PHI) from the CDW. These queries may be based on saved queries from i2b2 with associated more detailed information beyond the parameters allowed for the de-identified data set. Investigators have to go through a data request process with a biostatistics/honest broker consultation as well as IRB review. First, requests are reviewed by a Data Request and Review Committee (DRRC) to ensure that the requests comply with the DCA. This committee is made
up of one or more representatives from each of the HSSC institutions that have data in the HSSC CDW. The Honest Broker function is performed by individuals with operational access to the CDW as system custodians. All data sets requested must be, at a minimum, reviewed by the appropriate IRB, and may include certification as not-human-subject-research, or require exempt, expedited, or full board approval prior to the fulfillment of the data request.
3) Investigator-level Data Examples
Many MUSC investigators have data sets from clinical and population-based studies that may be useful for secondary analysis. Many principal investigators with long-standing extramural funding have archived data that can be used for secondary analyses and supplemental grant applications. For example, authors of this report have access to a national cohort study on stroke etiology including several supplemental studies, as well as several other cohort and case-control studies. Another example is the public use data set (PUDS) from the IMS III study of patients with acute ischemic stroke. This PUDS contain clinical, quality of life and economic data on 656 patients with acute ischemic stroke who were treated with inter-arterial or intra-venous tPA and followed for one year after stroke. A re-analysis of these data set could be used for pilot data for another study, or for student research projects. Numerous other PUDS are available, but faculty and students may not be aware of their existence or understand their content limitations.

4) Social Determinants of Health Database
**South Carolina Social Determinants of Health Database.** During Summer 2014, the Section of Health Systems Research & Policy within the Division of General Internal Medicine and Geriatrics will be creating the Geographical Social Determinant of Health database covering areas across SC. Anticipated variables include poverty rate, unemployment rate, education attainment, rate of persons registered to vote, Air Quality Index, housing vacancy rates, primary care physicians per 100,000 residents, and mean travel time to work. Information will be retrieved at the zip code level and sources will include US Census, Dartmouth Health Atlas, US Bureau of Labor Statistics, and the Environmental Protection Agency.

5) South Carolina Medicaid Data
**The SC Medicaid Data** are released through the SC Office of Research and Statistics. These data sets contain all bills for Medicaid patient contact with the medical care system, and are called encounter data. Encounter-level data files contain individual patient-level data using encounter-level data elements; release of these files requires an application and a signed Data Use Agreement. However, the ORS has permission to release aggregate customized reports based on encounter-level data without a signed agreement. The files contain bills for inpatient, outpatient, office visits, therapy sessions, physician bills for procedures, diagnostic test bills as well as all prescription drug bills. An exhaustive list of the variables included is available in the Appendix. There will be a processing charge for the data sets. This may range from $1,000-$2,000 or more depending of the volume of the requested data. The time horizon for gaining access is about 3-4 months because of the permission process. The data use agreement (DUA) requires original notarized signatures by a Dean, Informatics Director, and well as by the PI.