



HHS Public Access

Author manuscript

Int J Healthc Inf Syst Inform. Author manuscript; available in PMC 2020 September 09.

Published in final edited form as:

Int J Healthc Inf Syst Inform. 2018 ; 13(2): 1–12. doi:10.4018/ijhisi.2018040101.

Utilizing Combined Claims and Clinical Datasets for Research Among Potential Cases of Rare Diseases

Kevin J. Bennett,

University of South Carolina, Columbia, SC, USA

Joshua Mann,

University of Mississippi Medical Center, Jackson, MS, USA

Lijing Ouyang

Centers for Disease Control & Prevention, Atlanta, GA, USA

Abstract

With data quality issues with administrative claims and medically derived datasets, a dataset derived from a combination of sources may be more effective for research. The purposes of this article is to link an EMR-based data warehouse with state administrative data to study individuals with rare diseases; to describe and compare their characteristics; and to explore research with the data. These methods included subjects with diagnosis codes for one of three rare diseases from the years 2009–2014; Spina Bifida, Muscular Dystrophy, and Fragile X Syndrome. The results from the combined data provides additional information that each dataset, by itself, would not contain. The simultaneous examination of data such as race/ethnicity, physician and other outpatient visit data, charges and payments, and overall utilization was possible in the combined dataset. It is also discussed that combining such datasets can be a useful tool for the study of populations with rare diseases.

Keywords

Health Information Technology; Persons with Disability

INTRODUCTION

The growth in the use of health information technology (HIT) has been substantial in the United States of America. The percentage of office-based physicians who utilized some form of an electronic health record (EHR) increased from 18% in 2001 to 83% in 2014 (Hsiao & Hing, 2014). Likewise, 96% of US hospitals had an electronic health record system as of 2015, although this percentage was lower for smaller or rural hospitals (68 & 70%, respectively) (Hsiao & Hing, 2014). HIT utilization has also grown across health care sectors, such as in emergency departments (Selck & Decker, 2016) and home health

Publisher's Disclaimer: DISCLAIMER

Publisher's Disclaimer: The findings in this manuscript are those of the authors and do not necessarily represent the views of the South Carolina Department of Health and Human Services, the South Carolina Public Employee Benefit Authority, the South Carolina Data Oversight Council or the South Carolina Revenue and Fiscal Affairs Office.

agencies (Mitchell, Bennett, & Probst, 2013). Much of this growth was spurred by the 2009 Health Information Technology for Economic and Clinical Health (HITECH) Act, which gave financial incentives to institutions that adopted electronic medical record (EMR) and other HIT technologies. The HITECH Act introduced meaningful use (MU), which established a set of criteria to ensure that the HIT adopted was clinically useful and eligible for the incentives (Blumenthal & Tavenner, 2010). Part of the MU criteria is to view, download, or transmit patient data electronically (health information exchange – HIE); as of 2015, 33% of providers and 14% of hospitals were able to do this (ONC, DHHS). In addition, up to 80% of hospitals were able to conduct some form of external electronic query of patient data (ONC, DHHS).

Health information also holds promise for health services researchers. By capturing patient, visit, and other data during encounters, researchers are able to get a fuller picture of health care utilization, outcomes, and the factors that affect them both. Data derived from clinically-based HIT systems is limited in its utility for researchers in several ways. First, data can be difficult to obtain unless the researcher has an established relationship with the HIT system owner (i.e., providers or organizations) and a method to retrieve the desired data. Additional difficulties can arise due to limitations or exclusions from the data to protect privacy (ensuring all data is adequately protected and/or de-identified), obtaining patient consent for the use of the data, or ensuring the completeness and appropriateness of the data extracted. Second, there are few HIT-based data sources that are representative of larger populations beyond service areas or insured populations. While larger hospital systems may have a larger service footprint, their data may still only capture their patient population and not be representative of an entire community, or may only capture one segment of care within the care continuum, such as inpatient care or ambulatory care. Third, HIT systems often have separate service lines, meaning that one department may not use the same HIT system or dataset as another department in that same health care system or network. While some institutions are working towards unified HIT systems and datasets (e.g., Veterans Administration, Kaiser Permanente), these are also limited to data from their own enrollees. Finally, data is often collected and catalogued in a manner that serves the clinical care needs of providers and delivery systems, and not the needs of researchers (Miriovsky, Shulman, & Abernethy, 2012; Rosenbloom et al., 2011).

For these reasons, the Healthcare Information and Management Systems Society includes in its definition of an electronic health record (EHR) the need to support ongoing performance measurement as well as support clinical research and evidence based research, facilitated by health information exchange (Handler et al., 2016). Larger databases that include clinical data derived from electronic health records have potential to provide sources of research data, while also providing an opportunity to shift the thinking of how research is conducted, towards a more rapid-based learning model (Abernethy et al., 2010; Etheredge, 2007). This model seeks to capitalize on the clinical content of the data, but the regular updating of such data allows for more time sensitive approaches to research.

There is abundant research performed using a variety of data sources, including insurance claims data, surveys, and vital records (i.e., birth and death certificates). These data are utilized to examine a variety of topics, and have resulted in a wide range of findings

(Büchele et al., 2016; Faurot et al., 2015; Lauer & McCallion, 2015; Martin, Osterman, & Sutton, 2010; Ross et al., 2000); however, they can be limited. Claims data are useful repositories of individual and organizational information, but often lack key patient and clinical information (such as comorbidities, severity of the diagnoses, lab values, or patient preferences). Other data sources may include more in-depth patient information, but may be based upon patient self-report of conditions, utilization, or health outcomes and therefore may not be reliable (National Center for Health Statistics, 2015; Pierannunzi, Hu, & Balluz, 2013). The ability to utilize data drawn from clinical documentation would be helpful for improving the quality, depth, and explanatory power for health services research by supplementing these administrative sources.

In 2014, a project began that sought to utilize a combined dataset derived from a claims-based dataset and a medical record based dataset for the years 2009–2014. The purpose of this project, described below, was to determine how to link a clinical dataset with state administrative and claims datasets to analyze data on people with potential cases of rare diseases (e.g., Spina Bifida, Muscular Dystrophy, and Fragile X Syndrome). We then sought to describe and compare the characteristics of the individuals identified, and to explore potential research questions and activities that can be conducted using combined data.

METHODS

Dataset Creation

This particular project, funded by the Disability Research and Dissemination Center, focused on individuals with billing codes for one of three rare diseases: Spina Bifida (International Classification of Diseases version 9 (ICD9) 741.xx), Muscular Dystrophy (ICD9 359.0, 359.1, and 359.21), and Fragile X Syndrome (ICD9 759.83). These conditions were chosen for study as they are relatively rare in the general population, and are therefore not widely studied. Individuals were included in each dataset if they had at least one encounter with a diagnosis code including one of these codes. The data are limited, however, in the ability to confirm a medical diagnosis of such diseases without full medical chart review.

Claims and Administrative Data—The Revenue and Fiscal Affairs (RFA) Office¹ is a neutral service entity that warehouses various administrative data from state agencies, which is able to link persons across multiple service providers, locations, and data sources. Through a series of statutes and agreements, state agencies and organizations grant the RFA access to data and data systems. For this project, the RFA provided two sources of data; the first was from Medicaid and State Health Plan Claims, and the second was from Hospital administrative datasets (all described below). The RFA has developed a series of algorithms using combinations of personal identifiers to create its own unique identifier, which enables analysts to link data by person across the multiple sources, while protecting confidentiality of the patients. All data requests were approved by all participating agencies and

¹See <http://rfa.sc.gov/healthcare>

organizations, including the South Carolina Data Oversight Council and the University of South Carolina Institutional Review Board.

The first dataset included claims for patients with a service coded with one of the above mentioned rare disease billing codes from State Health Plan (SHP) or Medicaid (fee for service) as their insurance provider (hereafter referred to as the SHP/Medicaid dataset).

The SHP is an insurance option, administered by Blue Cross / Blue Shield of South Carolina that is offered to state and local government employees and their dependents. The other dataset is based upon information from the UB-04 billing and HCFA-1500 claim forms, utilized by facilities to bill for their services. These data include diagnostic information (International Statistical Classification of Diseases version 9, ICD-9), types of services received (either ICD-9, Healthcare Common Procedure Coding System(HCPCS), or Current Procedural Terminology (CPT) codes), service dates (admission and/or discharge dates), patient disposition, and other visit information for all inpatient and emergency department visits in the state, regardless of payer (hereafter referred to as the all-payer dataset).

The RFA queried these two data sources for patients with the ICD9 codes as above, and provided two separate datasets (one for SHP/Medicaid, and the other for the all-payer claims data). Race/ethnicity is not included in the State Health Plan data, and is thus not included in the SHP/Medicaid dataset; other information included in this dataset were age group, gender, and information regarding encounters and medications filled.

Electronic Health Records Data—SC is also home to Health Sciences South Carolina (HSSC), which houses the Clinical Data Warehouse (CDW)². HSSC is a statewide research collaborative focused on improving the quality of healthcare delivery, improving health information systems, training the healthcare workforce, and improving patient outcomes. Partners include seven of the state's largest health systems and the state's largest research-intensive universities. This is among the first systems to bring together data from three major research universities and several large health care systems. HSSC created the CDW to facilitate these goals; the CDW aggregates de-identified electronic patient records from these systems to enable providers and researchers to access patient data for research or treatment purposes. The data collected are from electronic health records related to inpatient, outpatient, and emergency department encounters from the affiliated systems.

Using the ICD9 codes described above, individuals were identified as potentially having one of these conditions in the CDW dataset (2009–2014). Information extracted from this dataset included patient demographics (sex, race, ethnicity, and area of residence – rural or urban), visit information (date, visit type (emergency, inpatient, other)), payer (commercial, government, other, self-pay), admission and discharge information, and diagnoses associated with the visit.

HSSC then sent their data file to the analysts at the RFA to merge with the two state-based datasets. To further alleviate privacy concerns, the service dates for all data were shifted

²See <https://healthsciencessc.org/tools-services/tool/clinical-data-warehouse/219>

randomly, but by a fixed amount across an individual's encounters. The RFA then utilized a proprietary linkage algorithm to match patients from the three sources (CDW, Medicaid/ SHP, all-payer datasets). Each person with unduplicated data was assigned a random number generated by an algorithm, referred to as the Unique ID. The algorithm uses personal identifiers that include, but are not limited to: first name, middle initial, last name, date of birth, race, and gender. The dataset was edited (i.e., characters were removed from numeric fields, dates were compared to valid ranges) and standardized (i.e., all characters were converted to uppercase) before being run through the algorithm. All of the identifier fields on a record housed at RFA were compared to records in the CDW data to identify exact or partial matches of individuals. A 'MatchType' (a string of characters that represent the match for each identifier field) was then assigned for each potential match. A 'MatchScore' (the rate indicating how often this 'MatchType' yields the same person when compared across multiple rows of data) and a 'FieldScore' (the number indicating the number of the fields that matched) was then assigned to the record. For each individual original record for which there were records with a MatchScore of 90 or greater, a single record was chosen as the actual match based on a match with more specific identifiable fields, such as social security number and date of birth. Once that single record match was chosen, the unique ID number from the matched record was assigned. If no records for a given original record have a MatchScore of 90 or greater, candidates for that original record were then compared to other possible candidates with lower MatchScore values. The process of scoring was then repeated for all records in this pool. The algorithm accounts for misspelling, name changes, and slight differences in the date of birth. An individual's Unique ID remains associated with that individual on all subsequent episodes of services in data captured by RFA, regardless of data source or service provider. Those that had an initial MatchScore of at least 90 were considered strong matches. The remaining were then compared on other characteristics to find additional matches, with a lower limit MatchScore of 60.

The final merged dataset was created using the RFA-generated unique identifier. The demographic variables were then compared and reconciled. In cases of a discrepancy, the assignment was based upon one of the state based datasets (all-payer or SHP/Medicaid).

Health care encounters were matched based upon the patient's ID, the data of the visit, and the setting and categorized into five categories: Inpatient, Emergency Department, Outpatient, Office Visits, and Other. Each dataset included different taxonomies to label their encounters; each utilized Inpatient, Emergency department, and Outpatient while the SHP/Medicaid included both Office and Other categories. To avoid double counting a visit recorded in more than one data source, a stratified matching method was utilized. First, encounters with the same date, by ID, were identified. Next, depending upon the dataset, the category of service was identified (Inpatient, Outpatient, or Emergency Department). Those labeled as 'Office' encounters in the SHP/Medicaid were matched with encounters labelled as "Outpatient" in the CDW data when appropriate. Second, if the date and category matched on two encounters, this was determined to be one visit. Non-matches were designated as different encounters.

Merging the data is simple enough, if done by the unique identifier identified by the matching algorithm described above. In practice, a major hurdle to an accurate merge was in

identifying a single encounter as just one encounter, and not confusing the multiple claims associated with an encounter as separate events. For example, a single physician visit may result in multiple claims, associated with tests, labs, or consultations. Multiple claims were merged into a single event for the accuracy of counting encounters. Also, with the merge, the same event could be represented in each of the datasets; thus, care must be taken to match events by date of service, service type, and service location. In addition, professional claims were also added to applicable events (e.g., professional claims associated with an inpatient claim or outpatient surgery).

RESULTS

Description of the Data

The resulting dataset derived from the above process was comprised of 2,637 patients identified within the CDW; 3,074 within the SHP/Medicaid data; and 2,675 from the all-payer data, for a total of 5,308 unique patients. When merged, 8.4% were in both the CDW and SHP/Medicaid; 8.3% were in both the CDW and all-payer; 11.8% were in both the SHP/Medicaid and all-payer; and 14.7% were in all three. Thus, 56.7% of the subjects were found in only one of the three datasets (see Figure 1).

Table 1 displays the demographic characteristics of those in each dataset, compared to those in all three datasets. The all-payer data had a higher proportion of females (56.2%) than the other datasets; the combined data was comprised of 54.2% female. The all-payer data also had the highest proportion that were white (72.9%), compared to 68.1% for the CDW dataset and 54.5% for the combined data (SHP/Medicaid race data were suppressed). The all-payer data skewed older but the combined data skewed towards the 0–17 age group (39.1%).

Table 2 displays the demographic characteristics of the study population by disease. Of these, 61.2% had a code associated with Spina Bifida, 32.7% had a code associated with Muscular Dystrophy, and 6.1% had a code associated with Fragile X. Potential Spina Bifida patients were predominantly female, white, and under the age of 17. Potential Muscular dystrophy patients were predominately male, white, and under the age of 17. Potential Fragile X patients were also predominately male, white, and under the age of 17.

Utilization and Expenditures

Table 3 displays the summary of annual claims, by type and year, for the individuals in each dataset. For the combined data, 22.7% had one or more inpatient visits. This proportion varied by condition; a smaller proportion of those potentially with Fragile X had at least 1 inpatient stay (13.0%) compared to those potentially with Muscular Dystrophy (22.6%) and Spina Bifida (23.8%). This trend remained, with some variations, when annualized (see Table 3). For emergency department visits, 38.3% of the entire sample had one or more visits. This proportion also differed by condition; a smaller proportion of those potentially with Fragile X had one or more visit (26.2%) compared to those potentially with Muscular Dystrophy (36.5%) and Spina Bifida (40.5%). This trend also remained, with some variations, when annualized (see Table 3).

We next examined median annual charges, subset by category of service. For individuals with potential cases of any of the three conditions, combined inpatient visits had an annual median charge of \$25,410 and ED visits had an annual median charge of \$13,067. Inpatient charges were lower for potential Muscular Dystrophy and Fragile X patients, but higher for potential Spina Bifida patients. The ED charges were similar for Spina Bifida, slightly higher for Muscular Dystrophy, but lower for potential Fragile X patients.

An examination of the primary diagnoses for type of visit was also instructive (see Table 4). For ED visits, commonly cited diagnoses included upper respiratory infections, urinary tract disorders, digestive system disorders, and skin ailments. In addition to those listed above, several diagnoses were unique to the condition: for potential Spina Bifida patients, migraine headaches were the fourth most common primary diagnosis; for potential Muscular Dystrophy patients, schizophrenic disorders were the sixth most common; for potential Fragile X patients, epilepsy was the fourth most common.

For inpatient visits, commonly cited primary diagnoses included urinary tract disorders, live births, pneumonia, and complications from procedures. In addition to those listed above, several diagnoses were unique to the condition: for potential Spina Bifida patients, chronic skin ulcers were the second most common primary diagnosis; for potential Muscular Dystrophy patients, cellulitis and abscesses was the tenth most common; for potential Fragile X patients, Bulbus cordis anomalies and anomalies of cardiac septal closure were the most common primary diagnosis.

DISCUSSION

Using administrative claims and clinical data allowed for a more comprehensive analysis of the experiences of patients with potential cases of selected rare diseases. Each of the datasets used in this project has its relative strengths; the CDW is derived from clinical documentation, the SHP/Medicaid data represents all of the individual level claims for those enrolled in those programs, and the all-payer data represents data from all facilities. In this study, the study population nearly doubled in size due to this combination compared to using any of the datasets alone. In addition, the combined dataset allowed researchers to obtain more comprehensive information about a continuum of care for patients across facilities and provider types. This is particularly useful for those with potential cases of rare diseases, as more complete information about their utilization can now be analyzed.

The use of combined administrative and clinical datasets may have the potential to improve studies of other populations as well, in the following ways:

- Matching visit data across the datasets can help fill in 'missing' information (such as race/ethnicity, visit type, location, prescriptions, lab values);
- Combining facility and professional claims can allow researchers to obtain more complete information for inpatient facility encounters;
- Examining actual laboratory values, not just that the test was performed, might permit researchers to study the linkage between such values and outcomes and/or utilization;

- Expanding information regarding insurance type could lead to more nuanced examination by payer status;
- Providing a more detailed description of the service unit within an inpatient facility can generate a better understanding of the care delivery context for the researcher;
- Identifying comorbidities or complexities in care due to the addition of diagnosis codes and procedure codes from the merged data might improve ability of researcher to make connections;
- Examining the provider classification, such as specialty, associated with the claims could allow for a more comprehensive understanding of what providers these individuals are seeking care from;
- Examining charge data associated with the claim can create a better understanding of health care costs associated with such diseases together with cost-to-charge ratios;
- Utilizing a combined dataset may be helpful to policy makers who are tasked with monitoring health services programs, interventions, and funding as their decisions can be made based upon fuller and more complete information as well as better examinations of the outcomes of such decisions.

There are limitations to these data, however. The CDW data included a large amount of outpatient visit data, but the categorization was vague, making it difficult. Specifically, it is not always possible to differentiate outpatient encounters by type or location (e.g., outpatient surgery vs. ambulatory physician visit). This can be clarified by matching to the SHP/Medicaid visit data, but only for those individuals with information in both sources. If these outpatient claims could be more precisely identified, the utility of that data may improve. Issues regarding the quality of clinical data are not uncommon, and continue to be a barrier to using such data for research (Coorevits et al., 2013; Weiskopf & Weng, 2013).

There were also cases of contradictory information within matched individuals and claims. For example, approximately 1% of the demographic data did not match exactly (i.e., sex, age and race/ethnicity not matching upon merge). When such discrepancies occurred, we relied on data from a state-based source rather than the CDW data. There were also instances where potential matched claims had substantially different service unit categorizations; it was often unclear if these simply represented different claims on the same day, or a discrepancy in classification by the provider. These instances were treated, conservatively, as matches. In addition, there are various reasons why an individual would be presenting one or more of the databases, but not all there, such as seeking care in only one place of service not captured in all three datasets, not having insurance coverage, or even seeking care from out-of-state providers.

Finally, identifying potential cases using ICD9 coding is not perfect. It has been shown that accuracy of administrative data was low when using a straightforward criterion of a single ICD9 diagnosis code for muscular dystrophy. However, accuracy of finding true cases improved substantially with added information in the algorithm (Smith et al., 2017). Future

studies may combine medical chart review with administrative data to address the questions of ICD-9 codes validation for rare conditions so that these administrative data sources can be better used for research on rare conditions.

Despite these limitations and difficulties, such a merged dataset represents a useful tool for researchers and policy makers alike. The use of such data has the potential to lead to improvements in population health and service delivery in many aspects.

This analysis was supported by a cooperative agreement from the Centers for Disease Control & Prevention through the Disability Research and Dissemination Center. RFA-R14-002, EMR Rare Conditions. Rare Condition research using a combined Administrative / Clinical Data warehouse.

Biography

Kevin Bennett is an Associate Professor in the Department of Family & Preventive Medicine at the University of South Carolina School of Medicine, in Columbia, SC. He is also a Senior Research Associate at the South Carolina Rural Health Research Center. His research agenda focuses upon care delivery for vulnerable populations, such as non-whites, rural residents, and those with chronic diseases. His publications cover topics such as transitions of care across communities, facility availability and service provision, and finance of health care services. He has also analyzed health care systems and their process of care, related to health information technology, network development, and community outreach.

Joshua R. Mann, MD, MPH, was born in Greenville, Miss., and raised in Perkinston, Miss. He completed his undergraduate degree at Delta State University in 1992, and his Doctor of Medicine degree at the University of Mississippi Medical Center in 1996. He then completed a transitional year at Carraway Methodist Medical Center in Birmingham, Ala., followed by a residency in public health and general preventive medicine at the University of South Carolina School of Medicine, with a Master of Public Health degree from the University of South Carolina's Arnold School of Public Health. He served for three years as research director for the Medical Institute for Sexual Health in Austin, Texas, then joined the faculty at the University of South Carolina School of Medicine in 2002, where he served as medical director of Employee and Student Health until June 2015. He served as Preventive Medicine residency director from August 2005 to June 2015. In that capacity, he led the work group that created ACGME-required milestones for Public Health and General Preventive Medicine residents. Dr. Mann has also made a substantial contribution as a researcher, having been principal investigator or co-investigator on a number of research and other grants. He has authored almost 60 peer-reviewed articles, more than 25 as first author. His research interests include the health of people with disabilities, prenatal risk factors for childhood disabilities, epidemiology of mental health, and links between religion/spirituality and health. Dr. Mann joined the University of Mississippi Medical Center's faculty in July 2015 as professor and chair of the Department of Preventive Medicine.

Lijing Ouyang works at the National Center on Birth Defects and Developmental Disabilities, Centers for Disease Control and Prevention, Atlanta, Georgia.

REFERENCES

- Abernethy AP, Etheredge LM, Ganz PA, Wallace P, German RR, Neti C, & Murphy SB et al. (2010). Rapid-Learning System for Cancer Care. *Journal of Clinical Oncology*, 28(27), 4268–4274. doi:10.1200/JCO.2010.28.5478 [PubMed: 20585094]
- Blumenthal D, & Tavenner M (2010). The “Meaningful Use” Regulation for Electronic Health Records. *The New England Journal of Medicine*, 363(6), 501–504. doi:10.1056/NEJMp1006114 [PubMed: 20647183]
- Büchle G, Becker C, Cameron ID, Auer R, Rothenbacher D, & König HH (2016). Fracture risk in people with developmental disabilities: Results of a large claims data analysis. *Osteoporosis International*, 24, 1–7.
- Coorevits P, Sundgren M, Klein GO, Bahr A, Claerhout B, Daniel C, & Kalra D et al. (2013). Electronic health records: New opportunities for clinical research. *Journal of Internal Medicine*, 274(6), 547–560. doi:10.1111/joim.12119 [PubMed: 23952476]
- Etheredge LM (2007). A Rapid-Learning Health System. *Health Affairs*, 26(2), w107–w118. doi:10.1377/hlthaff.26.2.w107 [PubMed: 17259191]
- Faurot KR, Jonsson Funk M, Pate V, Brookhart MA, Patrick A, Hanson LC, & Stürmer T et al. (2015). Using claims data to predict dependency in activities of daily living as a proxy for frailty. *Pharmacoepidemiology and Drug Safety*, 24(1), 59–66. doi:10.1002/pds.3719 [PubMed: 25335470]
- Handler T, Holtmeier R, Metzger J, Overhage M, Taylor S, & Underwood C (2016). HIMSS electronic health record definitional model version 1.0. Retrieved from www.himss.org
- Hsiao CJ, & Hing E Use and Characteristics of Electronic Health Record Systems Among Office-based Physician Practices: United States, 2001–2013. (2014). NCHS Data Brief No. 143. Available from: <http://www.cdc.gov/nchs/data/databriefs/db143.pdf>
- Lauer E, & McCallion P (2015). Mortality of people with intellectual and developmental disabilities from select US state disability service systems and medical claims data. *Journal of Applied Research in Intellectual Disabilities*, 28(5), 394–405. doi:10.1111/jar.12191 [PubMed: 25994364]
- Martin JA, Osterman MJK, & Sutton PD (2010). Are preterm births on the decline in the United States? Recent data from the National Vital Statistics System. *NCHS Data Brief*, (39).
- Miriovsky BJ, Shulman LN, & Abernethy AP (2012). Importance of Health Information Technology, Electronic Health Records, and Continuously Aggregating Data to Comparative Effectiveness Research and Learning Health Care. *Journal of Clinical Oncology*, 30(34), 4243–4248. doi:10.1200/JCO.2012.42.8011 [PubMed: 23071233]
- Mitchell J, Bennett KJ, & Probst J (2013). Organizational Factors Associated with Health Information Technology Adoption and Utilization Among Home Health / Hospice Agencies. *International Journal of Healthcare Information Systems and Informatics*, 6(3), 46–59. doi:10.4018/jhisi.2011070104
- Rosenbloom ST, Denny JC, Xu H, Lorenzi N, Stead WW, & Johnson KB (2011). Data from clinical notes: A perspective on the tension between structure and flexible documentation. *Journal of the American Medical Informatics Association*, 18(2), 181–186. doi:10.1136/jamia.2010.007237 [PubMed: 21233086]
- Ross NA, Wolfson MC, Dunn JR, Berthelot J-M, Kaplan GA, & Lynch JW (2000). Relation between income inequality and mortality in Canada and in the United States: Cross sectional assessment using census data and vital statistics. *BMJ (Clinical Research Ed.)*, 320(7239), 898–902. doi:10.1136/bmj.320.7239.898
- Selck FW, & Decker SL (2016). Health Information Technology Adoption in the Emergency Department. *Health Serv Res.*, 51(1), 32–47. [PubMed: 25854423]
- Smith MG, Royer J, Mann JR, & McDermott S (2017, 1 12). Using Administrative Data to Ascertain True Cases of Muscular Dystrophy: Rare Disease Surveillance. *JMIR Public Health and Surveillance*, 3(1), e2. doi:10.2196/publichealth.6720 [PubMed: 28082256]
- The Office of the National Coordinator for Health Information Technology, U.S. Department of Health and Human Services. (2014). Electronic Health Information Exchange Performance Reported to the Medicare EHR Incentive Program. Retrieved from <http://dashboard.healthit.gov/quickstats/pages/eligible-provider-electronic-hie-performance.php>

The Office of the National Coordinator for Health Information Technology, U.S. Department of Health and Human Services. (2016). Hospital Progress to Meaningful Use by Size, Type, and Urban/Rural Location. Retrieved from <http://dashboard.healthit.gov/quickstats/pages/FIG-Hospital-Progress-to-Meaningful-Use-by-size-practice-setting-area-type.php>

The Office of the National Coordinator for Health Information Technology, U.S. Department of Health and Human Services. (2016). Hospital Capability to Electronically Query. Retrieved from <http://dashboard.healthit.gov/quickstats/pages/FIG-Hospital-Electronic-Query-Capability.php>

Weiskopf NG, & Weng C (2013). Methods and dimensions of electronic health record data quality assessment: Enabling reuse for clinical research. *Journal of the American Medical Informatics Association*, 20(1), 144–151. doi:10.1136/amiajnl-2011-000681 [PubMed: 22733976]

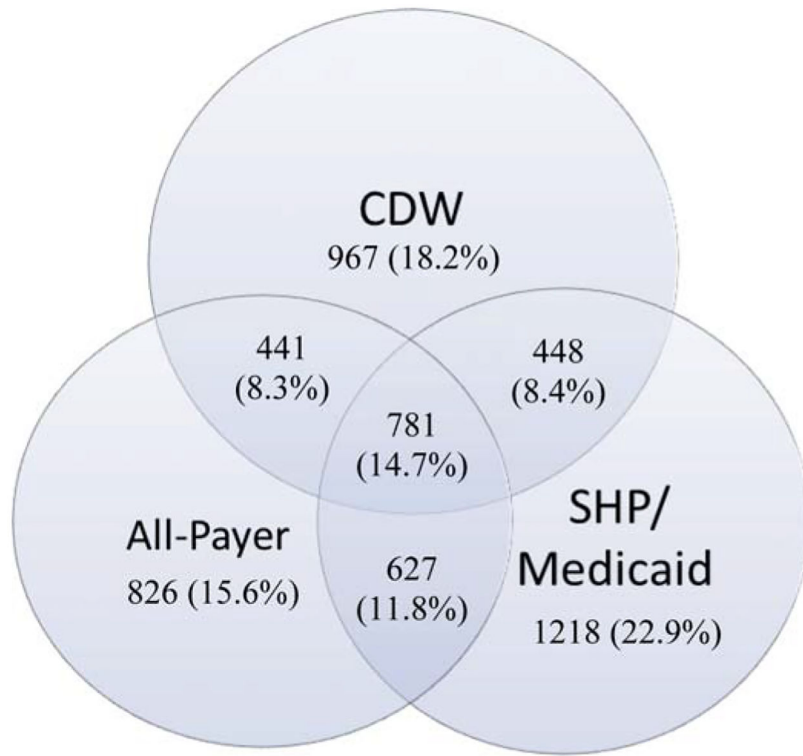


Figure 1.
Resulting dataset comprised of 2,637 patient identified within the CDW

Table 1.

Demographic characteristics of patient in administrative, clinical, and combined datasets; South Carolina, 2009–2014

| | CDW [†] | | SHP/Medicaid [†] | | All-Payer [†] | | All 3, Combined | |
|------------------------|------------------|------|---------------------------|------|------------------------|------|-----------------|------|
| | N | % | N | % | N | % | N | % |
| | 2637 | | 3074 | | 2675 | | 5308 | |
| Sex | | | | | | | | |
| Female | 1379 | 52.3 | 1608 | 52.3 | 1503 | 56.2 | 2875 | 54.2 |
| Male | 1212 | 46.0 | 1466 | 47.7 | 1172 | 43.8 | 2419 | 45.6 |
| Missing/Unknown | 46 | 1.7 | 0 | 0.0 | 0 | 0.0 | 14 | 0.3 |
| Race/Ethnicity* | | | | | | | | |
| White | 1796 | 68.1 | | | 1950 | 72.9 | 2895 | 54.5 |
| African American | 621 | 23.5 | | | 584 | 21.8 | 913 | 17.2 |
| Hispanic | 110 | 4.2 | | | 63 | 2.4 | 131 | 2.5 |
| Other | 110 | 4.2 | | | 78 | 2.9 | 151 | 2.8 |
| Missing/Unknown | 0 | 0.0 | | | 0 | 0.0 | 1218 | 22.9 |
| Age Group | | | | | | | | |
| 0–17 | 1016 | 38.5 | 1568 | 51.0 | 715 | 26.7 | 2078 | 39.1 |
| 18–24 | 248 | 9.4 | 283 | 9.2 | 224 | 8.4 | 417 | 7.9 |
| 25–44 | 618 | 23.4 | 796 | 25.9 | 863 | 32.3 | 1404 | 26.5 |
| 45–64 | 510 | 19.3 | 364 | 11.8 | 612 | 22.9 | 992 | 18.7 |
| 65+ | 245 | 9.3 | 63 | 2.0 | 261 | 9.8 | 417 | 7.9 |

* Race/ethnicity was suppressed from the SHP/Medicaid Dataset

[†] Proportions differ significantly, $\alpha = 0.05$, by data source

Table 2.

Demographic characteristic of patients potentially with rare diseases in combined administrative and clinical datasets; South Carolina, 2009–2014

| | Spina Bifida [†] | | Muscular Dystrophy [†] | | Fragile X [†] | | Any of the Three | |
|-----------------------|---------------------------|------|---------------------------------|------|------------------------|------|------------------|------|
| | N | % | N | % | N | % | N | % |
| All | 3252 | 61.2 | 1737 | 32.7 | 319 | 6.1 | 5308 | 100 |
| Sex | | | | | | | | |
| Female | 1990 | 61.2 | 795 | 45.8 | 90 | 28.2 | 2875 | 54.2 |
| Male | 1258 | 38.7 | 932 | 53.7 | 229 | 71.8 | 2419 | 45.6 |
| Missing/Unknown | 4 | 0.1 | 10 | 0.6 | 0 | 0.0 | 14 | 0.3 |
| Race/Ethnicity | | | | | | | | |
| White | 1707 | 52.5 | 1053 | 60.6 | 135 | 42.3 | 2895 | 54.5 |
| African American | 586 | 18.0 | 271 | 15.6 | 56 | 17.6 | 913 | 17.2 |
| Hispanic | 80 | 2.5 | 44 | 2.5 | 7 | 2.2 | 131 | 2.5 |
| Other | 91 | 2.8 | 45 | 2.6 | 15 | 4.7 | 151 | 2.8 |
| Missing/Unknown | 788 | 24.2 | 324 | 18.7 | 106 | 33.2 | 1218 | 22.9 |
| Age Group | | | | | | | | |
| 0–17 | 1382 | 42.5 | 505 | 29.1 | 191 | 59.9 | 2078 | 39.1 |
| 18–24 | 301 | 9.3 | 87 | 5.0 | 29 | 9.1 | 417 | 7.9 |
| 25–44 | 948 | 29.2 | 397 | 22.9 | 59 | 18.5 | 1404 | 26.5 |
| 45–64 | 482 | 14.8 | 482 | 27.7 | 28 | 8.8 | 992 | 18.7 |
| 65+ | 139 | 4.3 | 266 | 15.3 | 12 | 3.8 | 417 | 7.9 |

[†]Proportion differ significantly, $\alpha = 0.05$, by condition

Table 3.

Percent distribution of annual visits by patient potentially with rare diseases, by type of encounter, and year, in the combined administrative and clinical dataset; South Carolina, 2009–2014

| # of Visits | Spina Bifida [†] | | | Muscular Dystrophy [†] | | | Fragile X [†] | | | Any of the Three | | |
|-----------------------------|---------------------------|------|------|---------------------------------|------|------|------------------------|------------------|------|------------------|------|------|
| | 0 | 1 | >1 | 0 | 1 | >1 | 0 | 1 | >1 | 0 | 1 | >1 |
| Inpatient | | | | | | | | | | | | |
| 2009 | 77.3 | 14.4 | 8.3 | 78.1 | 13.0 | 8.8 | 88.8 | 5.2 | 6.0 | 78.3 | 13.4 | 8.3 |
| 2010 | 74.2 | 15.6 | 10.2 | 76.2 | 14.3 | 9.4 | 86.6 | 7.4 | 5.9 | 75.6 | 14.7 | 9.7 |
| 2011 | 74.0 | 14.8 | 11.2 | 73.6 | 14.7 | 11.7 | 83.7 | 10.0 | 6.3 | 74.5 | 14.5 | 11.1 |
| 2012 | 74.3 | 15.5 | 10.2 | 77.5 | 12.3 | 10.3 | 84.4 | 8.1 | 7.4 | 76.0 | 14.0 | 10.1 |
| 2013 | 75.6 | 14.8 | 9.6 | 76.0 | 13.5 | 10.6 | 86.7 | 6.8 | 6.5 | 76.4 | 13.9 | 9.8 |
| 2014 [*] | 83.9 | 11.7 | 4.4 | 85.6 | 10.3 | 4.0 | 92.6 | 7.4 [‡] | | 85.0 | 10.8 | 4.2 |
| Emergency Department | | | | | | | | | | | | |
| 2009 | 57.8 | 21.7 | 20.5 | 61.4 | 22.6 | 16.0 | 69.1 | 18.9 | 12.0 | 59.8 | 21.8 | 18.4 |
| 2010 | 56.1 | 22.8 | 21.1 | 61.5 | 20.7 | 17.7 | 75.5 | 13.0 | 11.5 | 59.2 | 21.5 | 19.4 |
| 2011 | 56.8 | 22.1 | 21.2 | 60.4 | 20.3 | 19.3 | 65.9 | 21.9 | 12.2 | 58.6 | 21.5 | 20.0 |
| 2012 | 56.4 | 21.5 | 22.1 | 60.2 | 20.6 | 19.3 | 71.5 | 15.6 | 13.0 | 58.6 | 20.8 | 20.6 |
| 2013 | 57.2 | 21.3 | 21.6 | 61.7 | 19.9 | 18.3 | 73.8 | 14.7 | 11.5 | 59.7 | 20.4 | 19.9 |
| 2014 [*] | 76.5 | 16.1 | 7.5 | 80.6 | 13.3 | 6.1 | 88.6 | 7.9 | 3.5 | 78.6 | 14.7 | 6.8 |

[†]Proportion differ significantly, $\alpha = 0.05$, by year and service category

[‡]Categories combined due to small sample sizes

^{*}Through August 2014

Table 4.

Ranking and frequency of diagnoses, by patients potentially with rare diseases, by type of encounter, in the combined administrative and clinical datasets; South Carolina, 2009–2014*

| | Spina Bifida (n) | Muscular Dystrophy (n) | Fragile X (n) |
|------------------------------|--|--|---|
| Emergency Department | | | |
| 1 | Chronic ulcer of skin (1103) | General symptoms (238) | General symptoms (36) |
| 2 | Symptoms involving head and neck (722) | Symptoms involving respiratory system / chest(215) | Acute upper respiratory infections of multiple sites (32) |
| 3 | Other disorders of urethra and urinary tract (682) | Other symptoms involving abdomen and pelvis (184) | Attention to artificial openings (29) |
| 4 | Migraine (646) | Other disorders of urethra and urinary tract (180) | Epilepsy and recurrent seizures (23) |
| 5 | Other symptoms involving abdomen / pelvis (497) | Symptoms involving digestive system (172) | Other congenital anomalies of upper alimentary tract (22) |
| 6 | Symptoms involving respiratory system / chest (457) | Schizophrenic disorders (140) | Diseases of hard tissues of teeth (20) |
| 7 | Other and unspecified disorders of back (446) | Other and unspecified disorders of back (126) | Other symptoms involving abdomen and pelvis (19) |
| 8 | General symptoms (403) | Diabetes mellitus (110) | Acute pharyngitis (16) |
| 9 | Symptoms involving digestive system (308) | Pneumonia, organism unspecified (109) | Diseases of esophagus (16) |
| 10 | Other disorders of bladder (339) | Chronic ulcer of skin (106) | Cellulitis and abscess of face (16) |
| Inpatient[†] | | | |
| 1 | Complications peculiar to certain specified procedures (212) | Care involving use of rehabilitation procedures (79) | |
| 2 | Chronic ulcer of skin (125) | Other diseases of lung (54) | |
| 3 | Other disorders of urethra and urinary tract (123) | Pneumonia, organism unspecified (52) | |
| 4 | Spina bifida (119) | Single liveborn (42) | |
| 5 | Other conditions of brain (100) | General symptoms (37) | |
| 6 | Single liveborn (98) | Muscular dystrophies and other myopathies (35) | |
| 7 | Other congenital anomalies of nervous system (84) | Other disorders of urethra and urinary tract (32) | |
| 8 | Other cellulitis and abscess (82) | Symptoms involving respiratory system / chest(28) | |
| 9 | Septicemia (73) | Complications peculiar to certain procedures (27) | |
| 10 | General symptoms (72) | Other cellulitis and abscess (26) | |

* Diagnostic categories using ICD-9 codes and the Clinical Classification Software (CCS) system developed by the Healthcare Cost and Utilization Project: <http://www.hcup-us.ahrq.gov/toolssoftware/ccs/ccs.jsp>.

[†] Inpatient diagnoses are not displayed for Fragile X due to small cell sizes, to ensure privacy