Let’s Get Technical: Hospital-Level Research Using Big Data

by Meghan Hufstader Gabriel, PhD, and Kendall Cortelyou-Ward, PhD

Abstract

As the field of health informatics and information management matures, students must be trained to evaluate not only the adoption rates of technologies but also the clinical, financial, and quality outcomes associated with these technologies. Big data can provide the necessary information to quantify these effects. This article provides a high-level overview of tools that faculty can use to instruct students on how big data is used to study hospital outcomes. Several data sources are discussed, and known issues in hospital-level research are detailed.

Keywords: big data; education; outcomes

Introduction

As the field of health informatics and information management matures, moving past implementation and into analytics and outcomes-based healthcare, students must be trained to take a holistic view of healthcare issues that may be presented to them in the real world. They must learn to evaluate not only the adoption rates of technologies but also the clinical, financial, and quality outcomes associated with the use of these technologies in large-scale healthcare systems. Big data can provide the necessary information to quantify these effects, but the data are commonly located in several data sets that are difficult to integrate, and accessing the data can be intimidating for students. In addition, knowledge of idiosyncrasies in the data sets can engender confidence and provide students with the skills required to understand the data.

This teaching and learning article provides a high-level overview of data resources that faculty can use to instruct students on data integration techniques for hospital-based research. The data sets discussed in this article include not only hospital characteristics but also financial reports, clinical data and outcomes, and technology use. However, the usefulness of these data sets cannot be realized until they are integrated with each other. Educators’ efforts to provide students with the needed skills in this area will assist graduates in the job market and ensure that accreditation standards are being met.

Accreditation Standards

As a sign of the changing field, curriculum competencies have evolved to include data-centered approaches to health informatics and information management. The recently revised criteria require that students at the graduate level, and to some extent students at the undergraduate level, demonstrate an understanding of data analytics and can use appropriate data tools to solve healthcare problems.[1]

AMIA Health Informatics Core Competencies

The American Medical Informatics Association (AMIA) developed a set of core informatics competencies that was approved by the board of directors in 2017. These competencies begin with foundational domains, which include Health Information Science and Technology and Social and Behavioral Science. Graduate students in health informatics are expected to demonstrate knowledge, skills, and attitudes reflecting these foundational domains and all of the content at their intersection. Of particular interest is the focus on Health Information Science and Technology, which challenges graduates to identify possible methods and tools to solve problems in health informatics.[2] These competencies were to be incorporated into academic curricula by August 2017.

AHIMA Graduate Health Informatics Curriculum Competencies

The American Health Information Management Association (AHIMA) also recently released a new set of curriculum guidelines for health informatics graduate programs. These competencies include nine domains covering a variety of core competencies, including Health Data Analytics. This domain specifies that graduates should be able to utilize tools to transform health data to improve decision-making and optimize health.[3] This focus on the ability to work with complex data indicates the importance of students' graduating with the skills necessary to work with complex data sets.

AHIMA Certified Professional in Health Informatics

AHIMA also recently announced the launch of a new credential in the health informatics field, Certified Professional in Health Informatics, which includes exam domains such as Data Analysis and Utilization, Data Reporting, and Data Management among the seven content areas. Individuals interested in pursuing this credential must have expertise in these areas to be successful in credentialing and in their career.

Problem Statement

As outcomes research comes to the forefront of the healthcare industry and becomes a necessary curricular component for accreditation, the importance of teaching relevant methods to perform this research has grown. However, the use of established data sets in the classroom can present a challenge for faculty. To be able to truly take a holistic view of a healthcare issue, one must integrate multiple data sets that may not have a linking variable. Knowing how to integrate these data sets is a crucial skill. This paper describes the most common data sets and their linking variables to provide a reference for faculty and students.

Hospital Identification Numbers/Linking Variables

Healthcare data sets commonly include one of two provider identification numbers used by the Centers for Medicare and Medicaid Services (CMS): the CMS Certification Number (CCN) and the National Provider Indicator (NPI). When these numbers are provided, they can often be used as linking variables to integrate data sets.

CMS Certification Number

The CCN is an identification number that is six digits in length, with the first two digits being identification codes for the state in which the provider is located and the last four digits identifying the type of facility.[4] This identifier has previously been known as the OSCAR Number, the Medicare Identification Number, and the Provider Number. For hospital researchers, numbers identifying the type of facility would fall between 0001 and 0879 for short-term hospitals and between 1300 and 1399 for

Critical Access Hospitals (CAHs). Although these numbers do not vary, the CCNs for specific hospitals can change because of closures, consolidations, and changes in CAH qualifications.

Researchers can use the CCN to explore trends in closures, consolidations, and CAH qualifications. The CCN identifies each Medicare provider and is used primarily to track provider agreements and cost reports. However, this identification number can be helpful to determine trends in hospital governance and status. The CCN serves a critical role in verifying that a provider has been certified by Medicare and identifying the type of services for which it has been certified. This number is used throughout the various components of CMS's business operations.

Hospital National Provider Indicator

The hospital NPI number is a 10-digit number that is used to identify healthcare service providers. It is national in scope and unique to the provider and is issued by the federal government through the National Plan and Provider Enumeration System. NPIs are assigned to individuals, groups, or organizations, including institutions such as hospitals. It is not uncommon for several NPIs to link to the same hospital. Unlike the CCN, in most cases a hospital's NPIs will not change. These numbers can be searched publicly and downloaded for free from the CMS website. However, the data set is extremely large and cannot be opened with ubiquitous platforms such as Microsoft Excel.[5]

Hospital-Level Data Sources

Hospital-level data sources include the Healthcare Information and Management Systems Society (HIMSS) data, the American Hospital Association (AHA) data, Hospital Compare, the Healthcare Cost and Utilization Project (HCUP), and the Provider of Services (POS) file. To conduct outcomes research using the multiple data sets available at the hospital level, it is important to be able to link the data sets together. Table 1, Figure 1, and Figure 2 provide insights into linking these data sets.

Healthcare Information and Management Systems Society Data

HIMSS Analytics conducts an annual study that collects information regarding approximately 5,500 nonfederal acute care hospitals in the United States. The data include information related to characteristics, technology, and staffing[6] and include a proprietary tracking algorithm assessing hospitals' progress toward a paperless patient record environment. Although the data set is proprietary, the Dorenfest Institute for Health Information,[7] part of the HIMSS Foundation, is an online resource that provides free data and reports regarding the use of technology in hospitals and information regarding characteristics of hospital systems from 1986 onward. These data are accessible free of charge under a university or government license.

The HIMSS data include several unique identifiers. The "UniqueID" variable is best for linking data because this number stays constant throughout the years in which these data are collected and therefore can be used for pooling information across years. The HIMSS identification variable "HAentityID" is unique only within each survey year and therefore cannot be used in longitudinal analysis (see Figure 3). The CCN is also included in the data set as the "MedicareNumber" variable; however, as noted above, the CCN for specific hospitals may change over time because of closures, consolidations, and changes in CAH qualification. The CCN allows the data set to be linked with other hospital-level data sources. If a matching identifier cannot be found, the hospital name and address provided in the data can be used to join the data sources.

American Hospital Association

The AHA Annual Survey Database is a proprietary source of hospital data including information on approximately 6,500 hospitals' finances, population health efforts, and health information technology.[8] For historical research purposes, it is critical to note that the AHA dates the survey by the year the survey was designed. Thus, the survey name is one year behind the date that responses were received by the hospitals. For example, the IT supplement data that were collected in 2011 are called the 2010 IT survey. However, this nomenclature changed as of the 2012 survey and name now reflects the year that the survey data was collected.

The identification number for these hospitals is called the “AHAID” and can be linked across years within all the data, including the supplementary survey data files. These data do not include other identifiers that could assist in linking this data set with other data sources. However, the HCUP provides a linkage file to assist researchers.[9] Therefore, depending on the data being merged, a match via hospital name or a crosswalk from the AHA data to another identifier will be necessary.

Hospital Compare

The Hospital Compare data sets are consumer-oriented data sets that include information regarding quality of care and hospital performance for more than 4,000 Medicare-certified hospitals in the United States.[10] These data are free to the public and provide facility-level information such as hospital characteristics, inspection information, the ability to receive lab data electronically, 30-day readmission rates, surgical complication rates, and certain patient satisfaction measures. The data include the hospital name, city, state, and latitude and longitude, which are useful for linking and mapping, as well as the CCN, which is referred to as “Provider_ID” in the data, and the hospital National Provider Identifier (NPI). These two identification variables allow for ease of linkage between data sets. These data are updated quarterly.

Healthcare Cost and Utilization Project

The HCUP provides State Inpatient Databases (SID), which can be used to identify, track, and analyze national trends in healthcare utilization, access, charges, quality, and outcomes.[11] The state-specific files contain information on all inpatient care records in the participating states, including diagnoses and procedures, admission and discharge status, patient demographics, payment source, charges, and length of stay. Currently, the 48 states that participate comprise more than 95 percent of all community hospital discharges in the United States.[12] The SID files encompass all patients, regardless of payer.

The HCUP data set includes not only its own schema for identification of hospitals (“HOSPID”) and each hospital’s NPI but also the “AHAID” data element for ease of linkage to the AHA data. The inclusion of these identifiers facilitates matching, allowing researchers to merge hospital characteristics and other information into their analytic files. This matching is especially useful for outcomes research. Although approximately 20 percent of AHA identifiers are missing from the HCUP data,[13] researchers can use hospital names, cities, and zip codes or other unique variables, such as the CCN, to link data with missing identifiers. These data are available for purchase from HCUP.

Provider of Services File

The POS file contains information on characteristics of hospitals and other facilities for each year, starting in 2006.[14] The data contain facility-level information for each Medicare-approved hospital. Previously, these data were available to the public for a fee, but they are now freely available for download. These data can also be linked by hospital name, city, and state. The data set includes the CCN variable, listed as “PRVDR_NUM.”

Known Issues with Hospital-Level Data

Because of hospital closures/openings, mergers/spinoffs, and changes in hospital structure, it is possible that hospital-level data, both public and proprietary, may have lags in updates that can result in challenges for data linkage. This concern is critical, especially when one considers hospital mergers. Hospital mergers have become commonplace in the US healthcare system, with nearly 900 hospitals having merged since 2000.[15] In addition, should a hospital change its structure—for example, changing from a small rural hospital to a hospital with CAH designation—the CCN would also change, resulting in data linkage issues. In the event of these changes, the American Hospital Directory (AHD) search function[16] may provide helpful information, including the hospital's CCN, name, city, and teaching status. The AHD also provides high-level information regarding closures, hospital structure changes, and merger rationales. Nonsubscribers have limitations on its use.

Discussion

The information presented in this article provides a starting point for faculty interested in integrating outcomes research into their curriculum and for researchers looking to use secondary data to conduct analysis of hospital-level data. As the field of healthcare becomes more data driven, it will be necessary for students to be able to link and analyze big data. Regardless of what informatics area students are focusing on, learning to analyze data is a critical step in developing the ability to help solve current challenges in the field. As our healthcare system searches for ways to optimize resource use and emphasizes value-based care, students will need to be well versed in outcomes research methods and big-data analytics in order to demonstrate the value of health information technology and management and improve associated healthcare outcomes.

Meghan Hufstader Gabriel, PhD, is an assistant professor in the Department of Health Management and Informatics at the University of Central Florida in Orlando, FL.

Kendall Cortelyou-Ward, PhD, is a program chair and associate professor in the Department of Health Management and Informatics at the University of Central Florida in Orlando, FL.

Notes

1. Commission on Accreditation for Health Informatics and Information Management Education (CAHIIM). *2017 Standards for Accreditation of Master's Degree Programs in Health Informatics (MHI).* 2017. Available at http://www.cahiim.org/documents/2017Standards_%20public_%20comment_%20watermark.pdf.
2. AMIA Health Informatics Core Competencies. 2017 Available at https://www.amia.org/sites/default/files/AMIA-Health-Informatics-Core-Competencies-for-CAHIIM.PDF.
3. AHIMA Graduate Health Informatics Curriculum Competencies. 2017. Available at https://www.ahima.org/~/media/AHIMA/Files/HIM-Trends/Draft%20AHIMA%20Graduate%20Health%20Informatics%20Curriculum%20Competencies_Completed.ashx?la=en.
4. Centers for Medicare and Medicaid Services. *Pub 100-07 State Operations Provider Certification.* October 12, 2007. Available at https://www.cms.gov/Regulations-and-Guidance/Guidance/Transmittals/downloads/R29SOMA.pdf.
5. Centers for Medicare and Medicaid Services. "NPI Files." Available at http://download.cms.gov/nppes/NPI_Files.html (accessed August 4, 2017).
6. HIMSS Analytics. *HIMSS Analytics Annual Study*. Available at https://www.himssanalytics.org/sites/default/files/Provider%20SS%20FINALTOPRINT.pdf.
7. HIMSS Foundation. "*The Dorenfest Institute for Health Information*." Available at https://apps.himss.org/foundation/histdata.asp (accessed August 7, 2017).
8. American Hospital Association. "AHA Data and Directories." Available at http://www.aha.org/research/rc/stat-studies/data-and-directories.shtml (accessed August 7, 2017).
9. Healthcare Cost and Utilization Project. "American Hospital Association Linkage Files." Available at https://www.hcup-us.ahrq.gov/db/state/ahalinkage/aha_linkage.jsp (accessed August 7, 2017).
10. Centers for Medicare and Medicaid Services. "About Hospital Compare Data." Available at https://www.medicare.gov/hospitalcompare/Data/About.html (accessed August 7, 2017).
11. Healthcare Cost and Utilization Project. "SID Database Documentation." Available at https://www.hcup-us.ahrq.gov/db/state/siddbdocumentation.jsp (accessed August 7, 2017).
12. Healthcare Cost and Utilization Project. "Overview of the State Inpatient Databases (SID)." Available at https://www.hcup-us.ahrq.gov/sidoverview.jsp (accessed August 7, 2017).
13. Healthcare Cost and Utilization Project. *HCUP Hospital Identifiers*. November 16, 2001. Available at https://www.hcup-us.ahrq.gov/db/maphosp.pdf.
14. Centers for Medicare and Medicaid Services. "Provider of Services Current Files." Available at https://www.cms.gov/Research-Statistics-Data-and-Systems/Downloadable-Public-Use-Files/Provider-of-Services/ (accessed August 7, 2017).
15. Schmitt, M. "Do Hospital Mergers Reduce Costs?" *Journal of Health Economics* 52 (2017): 74–94.
16. American Hospital Directory. Available at https://www.ahd.com/ (accessed August 7, 2017).

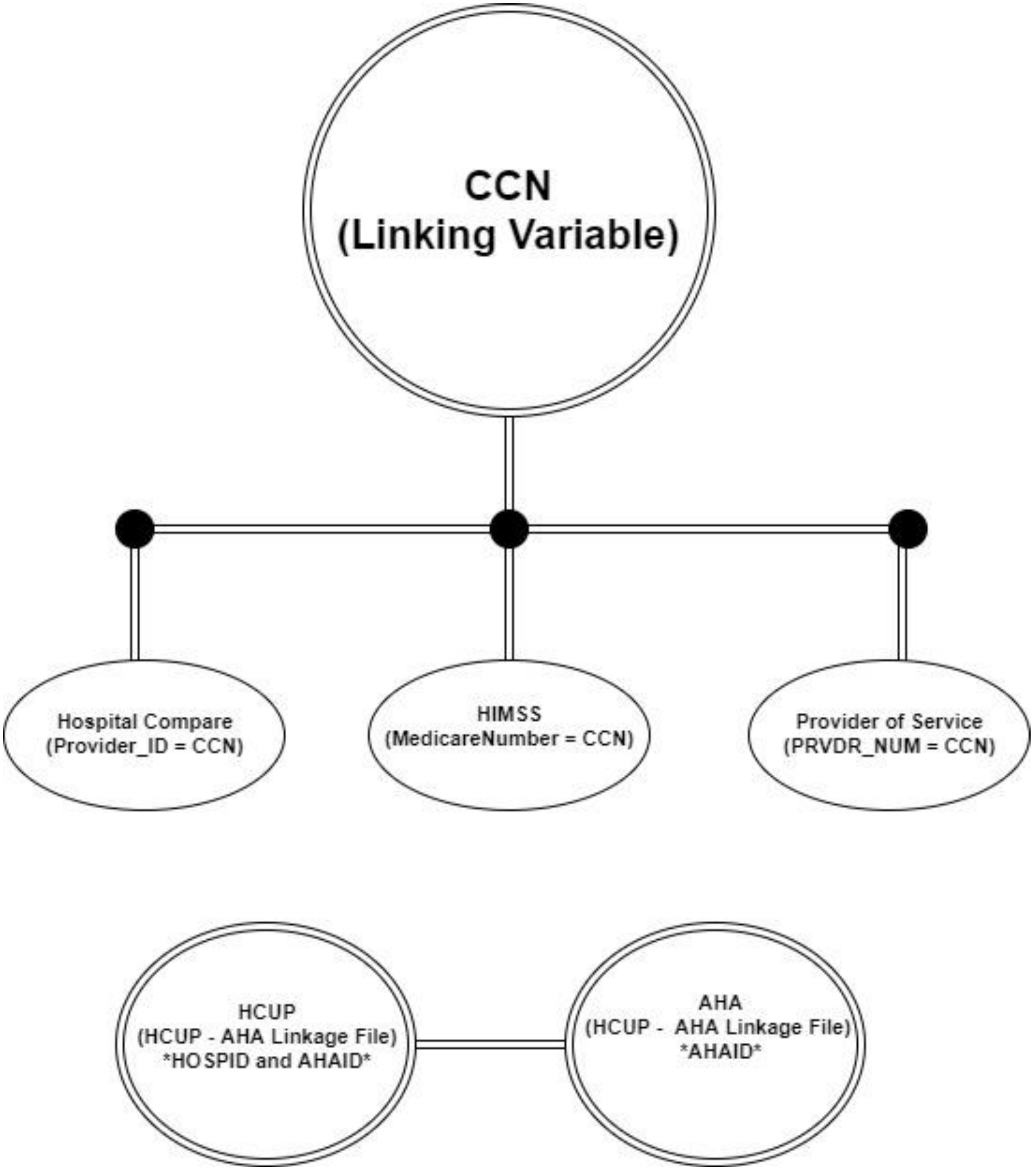
Table 1

Hospital-Level Data Sources for Health Information Management Research

Data	Unique Identifier	Linking Variable?	Sample	Procurement	Note
Health Information Management Systems Society (HIMSS)	UniqueID/HAentityID	Medicare Number (CCN)	5,500 hospitals and health systems in the United States	Proprietary, but the Dorenfest Institute for Health Information offers free data from 1986 onward for students	"UniqueID" does not change across years. "HAentityID" is unique only within each survey year.
American Hospital Association (AHA)	AHAID	None	6,500 nonfederal acute care hospitals in the United States	Proprietary; academic licenses are available. Subsets of the data may be available through research data sources such as the University of Pennsylvania's Wharton Research Data Services (WRDS).	Links to HCUP through the HCUP AHA linkage file
Hospital Compare	Provider_ID	CCN/Hospital National Provider Identifier (NPI)	4,000 Medicare-certified hospitals	Public	Provider_ID is the CCN.
Healthcare Cost and Utilization Project (HCUP) State Inpatient Databases (SID)	HOSPID	AHAID/Hospital NPI	95% of discharges from community hospitals in the United States (must combine to be at the hospital level)	Available for purchase from the HCUP central distributor	
Provider of Services	PRVDR_NUM	CCN	All Medicare-approved hospitals	Public	PRVDR_NUM is the CCN.

Figure 1

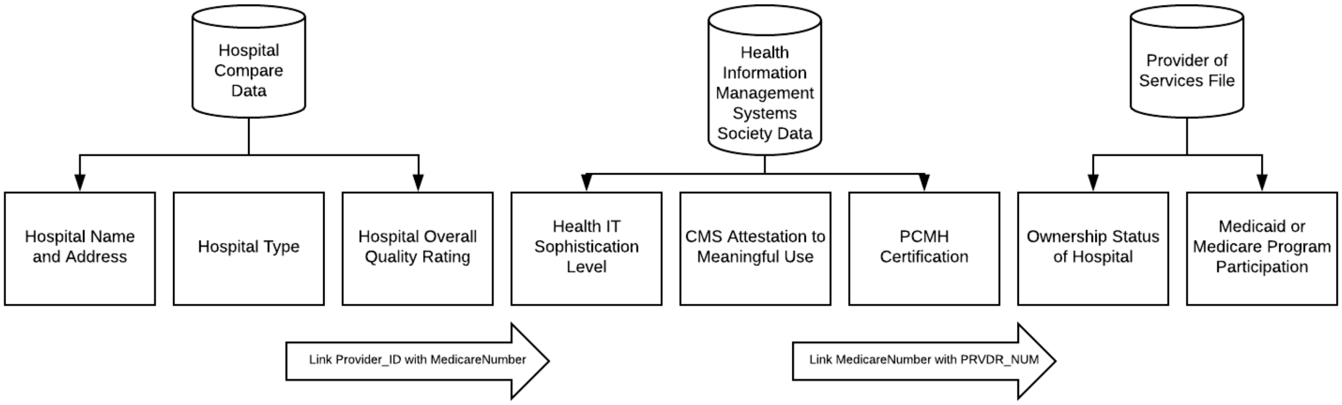
Framework to Link Variables across Hospital-Level Data Sets



Abbreviations: AHA, American Hospital Association; CCN, Centers for Medicare and Medicaid Services Certification Number; HCUP, Healthcare Cost and Utilization Project; HIMSS, Health Information Management Systems Society.

Figure 2

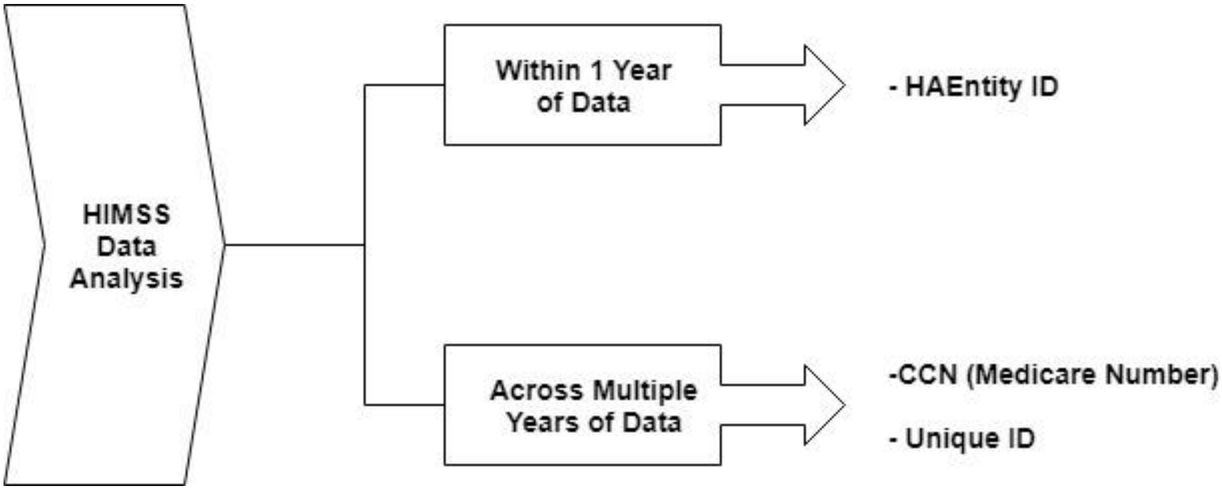
Example of Hospital-Level Information Linked across Data Sets


Hospital Compare Data
Health Information Management Systems Society Data
Provider of Services File
Hospital Name and Address
Hospital Type
Hospital Overall Quality Rating
Health IT Sophistication Level
CMS Attestation to Meaningful Use
PCMH Certification
Ownership Status of Hospital
Medicaid or Medicare Program Participation
Link Provider_ID with MedicareNumber
Link MedicareNumber with PRVDR_NUM

Abbreviations: CMS, Centers for Medicare and Medicaid Services; IT, information technology; PCMH, Patient-Centered Medical Home.

Figure 3

Health Information Management Systems Society (HIMSS) Data Analysis Framework, Within Years and Across Years



Abbreviation: CCN, Centers for Medicare and Medicaid Services Certification Number.