Aim 1: Design a database structured to allow a user to query any of the CMS-PPS datasets from 2012 through 2020

The initial goal of this aim was to allow for the query of only 2015-2020. This time range was chosen because of the ICD-9 to ICD-10 transition happened on October 1st of 2015 and I wanted to avoid the transition from ICD-9 to ICD-10 because it introduced mapping complexities that would be difficult to overcome. Ultimately, I concluded that I needed to add extra years to this dataset because not enough papers were identified in Aim 2 if data before 2015 was excluded. In total only two papers could be identified that fell under the original literature search described in Aim 2 of this paper and as a result, the scope of Aim 1 needed to be expanded in order to allow for more replications to be possible.

Methods

The Initial State of Aim 1

A Brief History of The OSU Secondary Datacore (SDC)

The SDC’s first incarnation occurred at The Center for the Advancement of Team Science, Analytics, and Systems Thinking in Health Services and Implementation Science Research (CATALYST). At CATALYST, we identified that multiple groups were purchasing the same datasets or different versions of a particular dataset. Furthermore, we found that researchers did not know what data existed, what data they needed to answer a particular research question, and how much that data would cost. There was no central repository for secondary medical datasets. SDC was our solution to that and I set out to build SDC as that central repository.

Initially SDC would contain three groups of datasets.

1 The American Hospital Association (AHA) Annual Datasets

2 The Healthcare Cost and Utilization Project (HCUP)

3 The CMS Claims Dataset

The AHA Annual datasets consist of two datasets that are released annually. These datasets are the AHA Annual Survey (AHAAS) and the AHA Information Technology Supplement (AHAIT). The AHAAS dataset is an annual survey completed by nearly 6,300 hospitals and more than 400 health care systems [AHA]. This dataset is one of the sources of hospital demographic and financial data and it is commonly used to identify hospital attributes. The AHAIT dataset contains questions that are particular to electronic medical records (EMR) and EMR related interfaces (how EMRs interact with other computer systems). The AHAIT is a dataset commonly used to discuss EMR complexity and capability. These datasets are aggregated at the AHA Membership level and therefore are typically used to investigate things at the hospital and health system level.

HCUP consists of five national level and three state level datasets that are released on an annualized basis by the Agency for Healthcare Research and Quality (AHRQ). The five state level datasets are the ones that are contained in the SDC. These datasets are the largest collection of longitudinal hospital care data in the United States [HCUP] and are commonly used to investigate trends in hospitalizations on a large scale. The five national HCUP datasets are the National Inpatient Sample (NIS), the Kids Inpatient Database (KID), the Nationwide Ambulatory Surgery Sample (NASS), the Nationwide Emergency Department Sample (NEDS), and the Nationwide Readmission Database (NRD). These datasets contain aggregated at the discharge level and therefore are used to investigate things aggregated by hospitalization.

The CMS Claims dataset… talk about this later.

The Heuristics of SDC Administration

1: Only use data and documentation from the purveyor.

At the start of the SDC, the first dataset that was put into the SDC was the AHA datasets. The biggest complication of loading the AHA datasets into the SDC was that, variables did not have consistent mapping between years. The AHA dataset we started with had questions labeled by their question ID in the survey and as a result I built a string-matching algorithm to identify similar questions between years. As it turned out, AHA had unique database identifiers for every question in their dataset and as a result the string-matching algorithm became unnecessary. The reason why this was not known initially was because the data that was used was not the data from AHA and did not have this database ID. So the first SDC heuristic was born- always start with data as it is provided by the purveyor.

2: If there is code that loads data into statistical software provided by the purveyor, use that code.

This rule came about as a result of loading HCUP into the SDC. AHRQ provides delimited flat files as well as SAS data files and a SAS load script with their dataset. When I first loaded HCUP datasets into the SDC I used the delimited flat files. As it turns out, the flat files do not match the AHRQ data dictionary and must be loaded into SAS in order to become the dataset as defined by AHRQ. So, as a result, the second heuristic was born- is the purveyor provides a load program, use it.

3: Automate everything that can be automated

In the beginning of this process, I was manually loading tables using the table loading wizard, in part, because I did not thing that I would need to be loading the same table over and over again. This was not the case. Use Python, SAS, or something as a controller when you design the data loads to perform the whole process of loading a dataset (or multiple datasets) into SQL at once.

The Initial State of CMS

The CMS data adventure began when we were added to the ACSC DUA 52416 and took over custodian ship. At that point I did not understand the CMS dataset

After Loading

How to get CMS SAF Data

SAF, RIF, CCW, RESDAC, CMS

Programming Languages and Tools

Dataset Initialization

Issues using the data

Solutions / Progenerated values

PQI / ELX / CCI

Next Hospitalization

APRDRG