This documentation is intended to describe the data used to calculate the values used for the visual presentations and numbers calculated as part of the Sword project.

The data dictionary is a tool used by software developers and data analysts to track and share the characteristics of the data they use and store in databases. It allows others in the organization to understand the origin and intent of values calculated from these data elements.

I am going to assume that if you made it to this paragraph, you will understand the value of this information and how to use it.

The cost reports used for Sword are provided in a formatted Excel spreadsheet on the website designed for this project - (kaloncon.com). Another file of all HHA cost reports for 2020 - 2022 is included as a CSV file.

Each of these data sources has the same fields or columns that include either data elements extracted from the HCRIS cost report data or calculated fields from this data.

As documented in my blog and the project Sword diary, there are many problems with the quality of cost report data. CMS and MedPAC use their own logic and process for deciding which cost reports to include in their calculations. This “cleaning” of the data is a normal process when developing data for analysis. The extent that this is required depends on the initial data quality and the specific quality issues associated with the data.

Although the cost report data as a whole is very “dirty”, the data we need is limited and is more available and reliable than most of the data from the cost reports. For all-payer margins we need total revenue and total expenses.

To be able to measure financial performance between groups of agencies, from year to year, and independently of the transition to MA, we need these financial totals associated with a consistent unit of measuring services. In the cost reports, we have a total annual census.

To explore the difference in the financial classes in the cost reports (Medicare, Medicaid, Other), we need accurate totals for each of these for revenue and census. Operating expenses can be calculated per census and compared to these revenue per census amounts to measure overall profitability by financial class and compare Medicare only margins to all-payer margins.

Finally, to analyze the impact of visits on profitability, we use visits by financial class. This requires reasonably accurate visit counts entered in the cost reports.

The fundamental flaw in the cost report data collection system is that each cost report value is entered by the provider or their vendor through a portal that has little to no data validation. You can enter revenue without census, totals that do match subtotals, and so on. You can even enter a report header without any values at all. These values are presumably entered by

someone using a document they have created for this purpose. Most are entered by vendors paid by providers to perform this task.

In some cases, I have had to make decisions regarding the data on what to do if two numbers that should match do not. This is the purpose of some of the calculated fields in this data. For example, the total of revenue for each of the three financial classes might not add up to the value entered for total revenue. When this is the case, I use the total that is the sum of the three financial classes. This maintains the proper ratio when these numbers are compared to each other as a ratio of the total.

Additionally, I have created rules that describe how I decided which cost reports to use in Sword and which ones to exclude. I describe this process, as I developed it, and details regarding these rules in the data validation portion of my diary. Here are the four rules. At least one of them applies to every cost report that was excluded from Sword.

- Eliminate all cost reports where Total Revenue or Total Census are 0.
- Eliminate all cost reports where one of the values for revenue or census by financial class is 0 and the other is not.
- Remove duplicated cost reports
- Eliminate all cost reports with a Z-Score greater than 0.03 and less than -0.03.

This is a list of both the spreadsheet and CSV fields or columns by the name in the header row in the order that they are represented in the data for each row or record. The values in bold come directly from the cost report data, the others are calculated for the Sword visualizations or for data validation. Where the descriptions are blank, the logic is repeated from the Medicare financial class.

A. **RPT_REC_NUM** - The report ID number assigned to the cost report
B. **FY** - Fiscal year
C. **MedicareRevenue** - Total annual Medicare Revenue for the FY
D. **MedicaidRevenue**
E. **OtherRevenue**
F. **TotalRevenue**
G. CalcTotalRevenue - Total of MedicareRevenue, MedicaidRevenue and OtherRevenue
H. **OperatingExpenses** - Total of allowed costs
I. **NetIncome** - Reported Net Income
J. CalcNetincome - CalcTotalRevenue - OperatingExpenses
K. **MedicareCensus** - Total annual Medicare census
L. **MedicaidCensus**
M. **OtherCensus**
N. **TotalCensus**
O. CalcTotalCensus - Total of MedicareCensus, MedicaidCensus and OtherCensus
P. CalcMargin - ((CalcTotalRevenue - OperatingExpenses) / CalcTotalRevenue)

Q. Z-Score - Statistical Z-Score based on the CalcMargin variance from the mean for each cost report
R. ExpensesPerCensus - OperatingExpenses / TotalCensus
S. MedicarePerCensus - MedicareRevenue / MedicareCensus
T. OtherPerCensus - OtherRevenue / OtherCensus
U. AllPerCensus - CalcTotalRevenue / CalcTotalCensus
V. MedicareMargin - ((MedicarePerCensus - ExpensesPerCensus) / MedicarePerCensus)
W. OtherMargin
X. AllMargin - ((AllPerCensus - ExpensesPerCensus) / AllPerCensus)

Other values are used from the cost report data and included in the Sword charts, but these are all that are required to calculate all-payer margins and margins by financial class.

If anyone has questions about this documentation in regards to recreating or validating it for this purpose, please contact me.