**Cost Reports Documentation**

By: Brendan Rabideau  
Created on: 4/18/17  
Updated on:   
Updated by:   
Purpose: To describe the process by which we take the raw CMS cost report data and use it to create provider-year cost estimation variables, such as cost-to-charge ratios. These variables can then be applied to Medicare claims data (where charge information exists) to estimate the cost of various health services.

**Raw Data:**

Begin by downloading the raw data from CMS in .CSV form. As of 2017, this data can be found here: <https://www.cms.gov/Research-Statistics-Data-and-Systems/Downloadable-Public-Use-Files/Cost-Reports/Cost-Reports-by-Fiscal-Year.html>

There are currently two versions of data – one is version 96 (going back to year 1996) and one is version 10 (starting in year 2010). The different versions represent a change in the organization of the cost reports where certain variables and worksheets were renamed, so different programs must be written to extract data after the switch. There are currently two related, but separate directories for these two versions. Place the appropriate dataset in the folders found here:

/schaeffer-b/sch-protected/from-projects/VERTICAL-INTEGRATION/rabideau/Data/Cost\_Reports/v96  
/schaeffer-b/sch-protected/from-projects/VERTICAL-INTEGRATION/rabideau/Data/Cost\_Reports/v10

Keep the same naming conventions used by CMS, as there are programs that will read in these files based on the current names. In v96 there will be 4 files – hosp\_YYYY\_NMRC, hosp\_YYYY\_RPT, hosp\_YYYY\_ALPHA, and hosp\_YYYY\_ROLLUP. In v10 there is no ROLLUP file.

* NMRC – Contains numeric information such as costs, charges, number of operating days, etc.
* RPT – Contains reporting information such as fiscal year begin and end dates
* ALPHA – Contains string variables such as provider ID, name and address
* ROLLUP – Appears to be a summary file with a combination of information from the above 3 files, but is only present in v96 and NBER claims it is unreliable

About the structure of the cost reports: The raw cost reports are long-form files that are very narrow. They are grouped by the variable rec\_num, which is a provider-number in a given year (so the same provider will have the same rec\_num in all cost reports in the same year, but will have a different rec\_num the next year). There is a variable for worksheet, line, and column number. The worksheet variable specifies which worksheet the observation is referencing, the line number says which line (row) on that worksheet is being referenced, and the column shows which column on that line on that worksheet is being selected. The final variable is the value of the particular cell that the worksheet-line-column combination points to in the cost reports. For example:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| rec\_num | wksht | line | col | value |
| 1 | D10A181 | 02700 | 0100 | 1000000 |
| 1 | D10A181 | 02800 | 0100 | 2000000 |
| 1 | D10A181 | 02900 | 0100 | 500000 |

We can see that rec\_num is the same in all 3 observations, so all of this information is attributable to the same provider. The worksheet D10A181 is one of the many sheets in the cost reports, and documentation on which one it is can be found at NBER (<http://www.nber.org/data/hcris.html>). In this case it points to the D1-Part 1 worksheet main provider for Medicare facilities. We can see that observation 1 references line 27, column 1. So we can look up that worksheet (currently found here <https://www.costreportdata.com/worksheet_formats.html>), and use those coordinates to locate the variable called “General inpatient routine service cost net of swing-bed cost (line 21 minus line 26)” and see that its value is 1,000,000.

Each rec\_num has thousands of observations that are identified in this way. The main goals of the programs described below are to:

1. Transform the cost reports into wide-form so that there is 1 observation per provider-year
2. Combine the different worksheets to aggregate the numeric, character, and reporting characteristics for each rec\_num
3. Identify the relevant variables to create cost variables, which will be Cost-to-Charge Ratio (CCR), Cost-Per-Day (CPD), Cost-Per-Visit (CPV), and Cost-to-Revenue Ratios (CRR) so that we can estimate costs in a given year for a given provider.

**Programs:**

There are separate, but similar programs for v96 and v10 files due to reorganization and renaming. The programs are currently found here:

* /schaeffer-b/sch-protected/from-projects/VERTICAL-INTEGRATION/rabideau/Programs/Costs/Test\_Old
* /schaeffer-b/sch-protected/from-projects/VERTICAL-INTEGRATION/rabideau/Programs/Costs/Reproduced

Additionally, there are different programs for different facility types broken into 3 categories: Hospital, free-standing SNF, and standalone HHA. The algorithms that extract data for all 3 facility types follows the same logic and each use 4 similar programs. The program are:

1. Inputhosp / inputsnf / inputhha:
   1. Reads in raw CSVs and makes .SAS7BDAT files for the hospital, SNF, or HHA cost reports. Notes that hospital-based SNFs are found in the hospital cost reports
2. Hosps2s3wkshts / snfs2s3wkshts / hhas2s3wkshts / hospsnfs2s3wkshts:
   1. Reads in the S2 and S3 worksheets from the ALPHA files. Flags relevant providers by provider type and creates the ‘universe of facilities’ for each facility type
3. Wide\_cost\_reports / snfccr / hha.standalone.cpv / hospsnfccr:
   1. Reads in relevant numeric data from the NMRC files, combines this with the S2S3 files (which are limited to only the facility type of interest at this point, e.g. only STACHs, only IRFs, etc.) and transposes this data so that all the information for a provider is on a single line. Then creates the appropriate cost variable (CCR, CRR, CPD, or CPV). In the case of HHA, this program also converts certain fiscal year values to calendar year values.
4. Combine\_trim\_hosp\_ccr2 / combine\_trim\_snf\_ccr / trim\_hha:
   1. For hospital and SNF facilities, these programs convert the relevant variables from fiscal year to calendar year. Then, for all facilities, a trimming procedure is applied to remove outlier values

**For Hospital Providers** (STACH, IRF, LTCH, and Hospital Based SNFs):  
NOTE: I will describe in detail the programs that generate our hospital measures that were summarized above. SNF and Home Health follow a very similar pattern, so for these facilities I will only go into detail about methods that diverge from the hospital measures.

Inputhosp.sas: Reads in the raw .CSV files and converts them into .sas7bdat files. This program is very straightforward and should be intuitive to update.

Hosps2s3wkshts.sas / hospsnfs2s3wkshts.sas: This program pulls in provider level data (ALPHA) from the S2 and S3 worksheets and flattens it out so that each rec\_num is its own line. In this program, variables are given the naming convention SX\_LLLLL\_CCCC, where SX can either be S2 or S3, LLLLL is the 5 digit line number (usually with leading or trailing 0’s), and CCCC is the 4 digit column number. Note that the column variable is 4 digits in v96, but 5 digits in v10 as this is essential for updating the programs in later years. We will use this naming convention later on as well. There are some special variables that have been given more detailed names. Among these variables are prov\_id, which is the provider number that can be used to link a facility across years, provider type, and the subprovider variables ho\_subX\_prov\_id, which are used to identify if a facility is the main provider (usually a hospital) or a provider nested in the same provider (like a hospital-based inpatient rehab facility). The subprovider variables are important for later. In addition, fiscal year begin and end dates are also merged on from RPT files. These variables will be used when converting the CCR from fiscal year to calendar year.

Wide\_cost\_report.sas / hospsnfccr.sas: Similar to the S2S3 program, one of the main purposes of this program is to take certain key worksheets containing numerical information (NMRC) and turn them from long-form to wide-form so that there is only 1 observation per rec\_num. Additionally, this program separates acute care hospitals, IRFs, long-term care hospitals, and hospital based SNFs. Finally, this program creates fiscal year CCRs in two different ways.

First, the program selects providers based on provider type using information from the S2S3 worksheets, created beforehand. Next, it takes the D1-Part1, D4, and C1 worksheets from the NMRC long-file, and creates a rec\_num level wide file, naming the variables with the convention “WWWW\_LLLLL\_CCCC” where WWWW is a 4-digit worksheet designation, and L’s and C’s are lines and columns. On the D1 and D4 sheets there is some nuance to the worksheet names related to subproviders.

For v96: The 3rd and 4th digits denote a supbrovider – 0A is a main provider (sub=0), 1B is the first subprovider (sub=1), 2B is the second subprovider (sub=2), and 0C is a hospital-based SNF

For v10: The 3rd and 4th digits denote a subprovider – 0A is a main provider (sub=0), 0B is an inpatient psychiatric facility subprovider, 0C is an IRF subprovider, and 0E is a hospital-based SNF.

The D1 and D4 worksheets contain information on costs and charges. D1 has general routine inpatient costs and charges, D4 has ancillary costs and charges. The other worksheet, C1, is a worksheet that also has general routine inpatient costs and charges, and we use C1 to fill in missing values from the D1 costs and charges worksheet because there are many missing fields in the D1 worksheet in v10. By summing the general inpatient costs and ancillary costs, and dividing by the sum of the general inpatient charges and ancillary charges, we get a yearly cost-to-charge ratio for each provider in the given fiscal year.

After merging the wide D1, D4, and C1 worksheets onto the already wide S2 and S3 worksheets of a particular facility type by rec\_num, we then create the two CCR variables – one using the D1 worksheet, and one using the C1 worksheet (called CCR\_c). The formulas for each type for v96 and v10 are as follows:

V96:

* Main Provider:
  + total\_CCR=((D1\_HOSP\_INPTN\_COST+D4\_HOSP\_ANCIL\_COST)/(D1\_HOSP\_INPTN\_CHARGE+D4\_HOSP\_ANCIL\_CHARGE));
  + total\_CCR\_c=((C1\_HOSP\_INPTN\_COST+D4\_HOSP\_ANCIL\_COST)/(C1\_HOSP\_INPTN\_CHARGE+D4\_HOSP\_ANCIL\_CHARGE));
* Subprovider1:
  + total\_CCR=((D1\_SUB\_INPTN\_COST+D4\_SUB\_ANCIL\_COST)/(D1\_SUB\_INPTN\_CHARGE+D4\_SUB\_ANCIL\_CHARGE));
  + total\_CCR\_c=((C1\_SUB\_INPTN\_COST+D4\_SUB\_ANCIL\_COST)/(C1\_SUB\_INPTN\_CHARGE+D4\_SUB\_ANCIL\_CHARGE));
* Subprovider2:
  + total\_CCR=((D1\_SUB2\_INPTN\_COST+D4\_SUB2\_ANCIL\_COST)/(D1\_SUB2\_INPTN\_CHARGE+D4\_SUB2\_ANCIL\_CHARGE));
  + total\_CCR\_c=((C1\_SUB2\_INPTN\_COST+D4\_SUB2\_ANCIL\_COST)/(C1\_SUB2\_INPTN\_CHARGE+D4\_SUB2\_ANCIL\_CHARGE));
* Hospital-Based SNF (different program, hospsnfccr.sas):
  + total\_CCR\_c= ((C1\_03400\_0500 + D4\_10100\_0300) / (C1\_03400\_0600 + D4\_10300\_0200));

V10:

* Main Provider:
  + total\_CCR=((D1\_HOSP\_INPTN\_COST+D3\_HOSP\_ANCIL\_COST)/(D1\_HOSP\_INPTN\_CHARGE+D3\_HOSP\_ANCIL\_CHARGE));
  + total\_CCR\_c=((C1\_HOSP\_INPTN\_COST+D3\_HOSP\_ANCIL\_COST)/(C1\_HOSP\_INPTN\_CHARGE+D3\_HOSP\_ANCIL\_CHARGE));
* Inpatient Psychiatric Facility:
  + total\_CCR=((D1\_IPF\_INPTN\_COST+D3\_IPF\_ANCIL\_COST)/(D1\_IPF\_INPTN\_CHARGE+D3\_IPF\_ANCIL\_CHARGE));
  + total\_CCR\_c=((C1\_IPF\_INPTN\_COST+D3\_IPF\_ANCIL\_COST)/(C1\_IPF\_INPTN\_CHARGE+D3\_IPF\_ANCIL\_CHARGE));
* Inpatient Rehabilitation Facility:
  + total\_CCR=((D1\_IRF\_INPTN\_COST+D3\_IRF\_ANCIL\_COST)/(D1\_IRF\_INPTN\_CHARGE+D3\_IRF\_ANCIL\_CHARGE));
  + total\_CCR\_c=((C1\_IRF\_INPTN\_COST+D3\_IRF\_ANCIL\_COST)/(C1\_IRF\_INPTN\_CHARGE+D3\_IRF\_ANCIL\_CHARGE));
* Hospital-Based SNF (different program, hospsnfccr.sas):
  + total\_CCR\_c= ((C1\_04400\_00500 + D4\_20000\_00300) / C1\_04400\_00600 + D4\_20000\_00200))

(See appendix for more detailed information about variable location in the cost reports)

Combine\_trim\_hosp\_ccr2.sas: This program, the final in the series, takes the fiscal year CCRs and converts them to calendar year CCRs to match our calendar year data. To do this, we use the fiscal year begin and end date variables (fy\_bgndt and fy\_enddt) to determine how many days in a given calendar year are reported in the given fiscal year for a facility. We then weight and sum the fiscal year CCRs by the reporting days in that calendar year. For example, a 2012 FY begin date of 10/1/2011 and FY end date of 9/30/2012 would mean that 273 days from this fiscal year were spent in CY 2012. Then FY begin date of 10/1/2012 and FY end date of 9/30/2013 would imply that 92 days of the 2013 FY were spent in 2012. Assuming the fiscal year CCR in 2012 is 0.5, and the fiscal year CCR in 2013 is 0.4, we would calculate the 2012 calendar year CCR as:

Cal\_CCR2012 = (273/365)\*0.5 + (92/365)\*0.4 = 0.4748

After calculating the calendar year CCR, this program also outputs a separate dataset with trimmed calendar year CCRs. The trimming procedure is as follows:

1. Restrict values to 0.1 < CCR <= 10, else CCR is set to missing
2. Calculate the mean and the standard deviation of the restricted CCRs in each year
3. If a calendar year CCR is more than 4 standard deviations away from the yearly mean for restricted CCRs, set this CCR to missing

The trimmed CCR guards against outrageous values that are clearly not within the realm of possibility for a cost-to-charge ratio, and does not noticeably reduce the number of facilities for which we report CCRs. The final output variables are called cal\_total\_CCR and cal\_total\_CCR\_c.

**For Freestanding SNFs:** The 1st, 2nd, and 4th stage are the same as above. There are minor differences in the 3rd step for calculating cost variables for freestanding SNFs because we do not calculate a cost-to-charge ratio, but rather a cost-to-revenue ratio and a cost-per-day variable. Therefore, we use different worksheets (described in the appendix) and calculate our cost variables as follows:

* snf\_cpd=SNF\_Total\_Cost/SNF\_Inpatient\_Days
* snf\_ccr=SNF\_Total\_Cost/SNF\_Inpatient\_Revenue
* total\_ccr=Provider\_Total\_Cost/Provider\_Total\_Revenue

(See appendix for variable locations in the cost reports)

**For Home Health Agencies:** HHAs have a modified version of the CCR to calculate costs called the cost-per-vist (CPV). In general, the home health programs differ from the hospital programs because there is only a single version – v96. The HHA cost reports were not updated in 2010 like the other facilities’ cost reports were. Barring this, the first two programs (inputhha.sas and hhas2s3wkshts.sas) are functionally the same as the hospital based programs. The third program, hha.standalone.cpv.sas, has significant differences. For HHA costs, we calculate a CPV variable using data from the HHA cost reports, whose worksheets are different from the hospital cost reports (worksheet formats and documentation can be found here <http://www.nber.org/data/hcris-hha.html>). The relevant data comes from worksheet C1, which provides yearly average costs per visit for a provider for 6 types of services – skilled nursing, physical therapy, occupational therapy, speech pathology, medical social service, and home health aide – as well as the S2 worksheet which provides the total number of visits using these services for a provider in a given year. We then create our composite cost-per-visit for a provider-year by multiplying the share of the total visits that utilize a given service by the average cost per visit of that service. We then sum the values for all 6 services. For example:

(0.1\*$100 + 0.2\*$200 + 0.3\*$300 + 0.4\*$100 + 0\*$0 + 0\*$0) = $140 per visit, for a provider that provided 10% SNF services, 20% PT, 30% OT, 40% speech pathology, and 0% other, with an average cost per service of $100 for SNF, $200 for PT, etc.

NOTE: The hha.standalone.cpv.sas program calls in a macro called ***MCALCONVERT*** which converts key values from the hha\_s2s3\_1996\_20XX dataset (produced by hhas2s3wkshts.sas) from fiscal year to calendar year. Among the important variables that get transformed are total number visits, Medicare number of visits, and other number of visits for each service type offered by home health agencies.

The trimming procedure is used on the calendar year CPV for each of the 6 services individually. The trimming rules are as follow:

1. Restrict values to 10<=CPV<8000, else CPV is set to missing
2. Calculate the mean and the standard deviation of the restricted CPVs in each year
3. If a calendar year CCR is more than 3 standard deviations away from the yearly mean for restricted service-specific CPV, set this CPV to missing

**COST REPORTS APPENDIX**

Worksheets and Key Variables for Hospitals (hosp\_cost\_report\_wksht\_v10.pdf):

* S-2 Part I – Hospital And Hospital Health Care Complex Identification Data (awksht= “S200001”)
  + Line 3 Column 2 – Hospital CCN (Provider ID)
  + Line 4 Column 2 – IPF CCN (Provider ID)
  + Line 5 Column 2 – IRF CCN (Provider ID)
  + Line 9 Column 2 – Hospital-Based SNF CCN (Provider ID)
  + Line 3 Column 4 – Hospital Provider Type
  + Line 4 Column 4 – IPF Provider type
  + Line 5 Column 4 – IRF Provider Type
* S-3 Part I – Hospital and Hospital Health Care Complex Statistical Data (awksht = “S300001”)
* D-1 Part I – Computation of Inpatient Operating Costs (nwksht= “D10A181” (Main provider), “D10B181” (IPF Subprovider), “D10C181” (IRF Subprovider), “D10E181” (Hospital-Based SNF))
  + NOTE: In the D1 and D3 Worksheets, there are not separate lines for subproviders like in the S2, S3, and C1 worksheets, there are sub-worksheets that use the same lines and columns instead. For example, IRF subprovider costs are worksheet D10**C**181, line 27, column 1; hospital-based SNF costs are worksheet D10**E**181 line 27, column 1
  + Line 27 Column 1 – General Inpatient Routine Service Cost Net of Swing-Bed Cost
  + Line 28 Column 1 - General Inpatient Routine Service Charges (excluding swing-bed and observation bed charges)
* D-3 – Inpatient Ancillary Service Cost Apportionment (nwksht= “D30A181” (Main provider), “D30B181” (IPF Subprovider), “D30C181” (IRF Subprovider), “D30E181” (Hospital-Based SNF))
  + NOTE: See D1 above. Additionally, D3 was previously D4 in the v96 version
  + Line 200 Column 2 – Total (Ancillary) Inpatient Program Charges
  + Line 200 Column 3 – Total (Ancillary) Inpatient Program Costs
* C Part 1 – Computation of Ratio of Costs to Charges (nwksht = “C000000”)
  + Line 30 Column 5 – Adults and Pediatrics (General Routine Care) Total Costs
  + Line 40 Column 5 – Subprovider IPF Total Costs
  + Line 41 Column 5 – Subprovider IRF Total Costs
  + Line 44 Column 5 – Hospital-Based SNF Inpatient Costs
  + Line 30 Column 6 – Adults and Pediatrics (General Routine Care) Inpatient Charges
  + Line 40 Column 6 – Subprovider IPF Inpatient Charges
  + Line 41 Column 6 – Subprovider IRF Inpatient Charges
  + Line 44 Column 5 – Hospital Based SNF Inpatient Charges

Worksheets and Key Variables for Freestanding SNFs (snf\_cost\_report\_all\_v10.pdf):

* S-2 Part I – Skilled Nursing Facility and Skilled Nursing Facility Health Care Complex Identification Data (awksht= “S200001”)
  + Line 4 Column 2 – SNF CCN (Provider ID)
* S-3 Part 1 – Skilled Nursing Facility and Skilled Nursing Facility Health Care Complex Statistical Data (awksht= “S300001”)
  + Line 1 Column 7 – SNF Total Inpatient Days
* B Part 1 – Cost Allocation – General Service Costs (nwksht= “B000001”)
  + Line 30 Column 18 – SNF Total Cost
  + Line 31 Column 18 – NH Total Cost
  + Line 100 Column 18 – Total Total Cost
* G-2 Part 1 – Statement of Patient Revenues and Operating Expenses (nwksht= “G200001”)
  + Line 1 Column 1 – SNF Inpatient Revenue
  + Line 2 Column 1 – NH Inpatient Revenue
  + Line 14 Column 3 – Total Total Revenue

Worksheets and Key Variables for HHA (hha\_cost\_report\_wksht\_all.pdf):

* S-2 – Home Health Agency Complex Identification Data (awksht= “S200000”)
  + Line 2 Column 2 – HHA CCN (Provider ID)
* S-3 Part 1 – Home Health Agency Statistical Data (awksht= “S300001”)
  + Line 1 Column 1 – SNF Medicare Visits
  + Line 1 Column 3 – SNF Other Visits
  + Line 1 Column 5 – SNF Total Visits
  + Line 2 Column 1 – Physical Therapy Medicare Visits
  + Line 2 Column 3 – Physical Therapy Other Visits
  + Line 2 Column 5 – Physical Therapy Total Visits
  + Line 3 Column 1 – Occupational Therapy Medicare Visits
  + Line 3 Column 3 – Occupational Therapy Other Visits
  + Line 3 Column 5 – Occupational Therapy Total Visits
  + Line 4 Column 1 – Speech Pathology Medicare Visits
  + Line 4 Column 3 – Speech Pathology Other Visits
  + Line 4 Column 5 – Speech Pathology Total Visits
  + Line 5 Column 1 – Medical Social Service Medicare Visits
  + Line 5 Column 3 – Medical Social Service Other Visits
  + Line 5 Column 5 – Medical Social Service Total Visits
  + Line 6 Column 1 – Home Health Aide Medicare Visits
  + Line 6 Column 3 – Home Health Aide Other Visits
  + Line 6 Column 5 – Home Health Aide Total Visits
* C Part 1 – Apportionment of Patient Service Costs
  + Line 1 Column 4 – SNF Average Cost Per Visit
  + Line 2 Column 4 – Physical Therapy Average Cost Per Visit
  + Line 3 Column 4 – Occupational Therapy Average Cost Per Visit
  + Line 4 Column 4 – Speech Pathology Average Cost Per Visit
  + Line 5 Column 4 – Medical Social Service Average Cost Per Visit
  + Line 6 Column 4 – Home Health Aide Average Cost Per Visit