

Discovering CMS Data for Healthcare Research

Laxminarayana Ganapathi, Matthew Urato, Jean Hastie, Merry Rabb
Center for Health Data Analytics, RTI International, RTP NC 27709

ABSTRACT

Center for Medicare and Medicaid (CMS) hosts a plethora of health care data and at the outset aims to disseminate the benefits of this vast information to the public, without compromising the individual's privacy. It is estimated that over 98% of adults age 65 and over are enrolled in Medicare and over 99% of the deaths in this group are accounted for in the Medicare program, thus making it a vital repository of data for healthcare research about the senior population of about 50 million. While the size of the data could be intimidating, researchers have taken up the challenge of working with this ocean of data to evaluate, model, hypothesize and help CMS in improving its services. It is welcome news for SAS users that most of the data is available as SAS data sets. This work aims to introduce the newcomers to the structure of the CMS databases and point to the need and advantage of a deeper understanding of the data for the success of a study.

INTRODUCTION

Several researchers have reviewed the uses of health care utilization databases for epidemiologic research on therapeutics. Schneeweiss and Avorn [1] have brought out the scope of these databases beyond the conventional gold standard, RCT (Randomized Clinical Trials). These researchers have discussed the limited scope of the traditional RCTs and the need go beyond to study real-world effectiveness and utilization patterns, and their availability at relatively low cost without long delays. Harpe [2] has discussed the use of secondary data Sources for pharmacoepidemiology and outcomes research. Verhamme, and Sturkenboom have suggested that pharmacoepidemiological studies are crucial in the research on the safety of drugs in children. They have concluded that knowledge on the different pharmacoepidemiology methods is important to guarantee optimal use and correct interpretation of the data [3].

Ever since government made it a goal to safely share digital data with public, several agencies have strived and HHS and CMS have made it their goal to, "Unleash the Power of Data and Innovation To Improve Health" [4]. There are websites dedicated to making high value health data more accessible to entrepreneurs, researchers, and policy makers in the hopes of better health outcomes for all. Wang et al have reviewed [5] the possibilities of a large database of longitudinal data like CMS and have discussed the standards for, 'Reporting to Improve Reproducibility', furthering the RECORD (REporting of Studies Conducted Using Observational Routinely-Collected Health Data) initiative [6]. There are other initiatives like STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) [7] and EQUATOR (Enhancing the QUALity and Transparency Of health Research) [8] to enhance the standards in reporting of the data. Even as the randomized clinical trials (RCT) [9] ruled as the gold standard for research to find causal relationships, large databases like CMS have been efficiently used to explore and test hypotheses using alternate designs like prospective cohort studies, retrospective cohort studies, and case control studies with propensity score matching. [10 -15].

OVERVIEW OF CMS DATABASE

Current CMS data is integrated from the administrative, transactional and medical records of the beneficiaries. Popularly they are also known by their components as Part A, B, C and D. It is estimated that in 2015, the national healthcare expenditure was 17.8% of the GDP [16]. Table 1 depicts the recent population figures from the CMS website. One can clearly appreciate the large number of the elderly and disabled population served by CMS. Table2 shows the latest financial data downloaded from the CMS website. A total of 1 trillion plus federal dollars are spent between Medicare and Medicaid services during fiscal 2016. Given these large figures, the government has large motivations to provide efficient care and reduce cost at the same time.

Table 1. CMS Program Data – Populations^a

Medicare (avg monthly)	CY 2015	CY 2016	CY 2017^b
Parts A and/or B	55.6	57.1	58.0
Aged	46.7	48.3	49.1
Disabled	8.9	8.9	8.9
Original Medicare Enrollment	37.8	38.4	38.0
MA & Other Health Plan Enrollment	17.8	18.7	20.0
MA Enrollment	16.5	17.6	18.6
Part D (MA PD+PDP)	39.5	41.2	42.5
Medicaid (avg monthly)^c	FY 2015	FY 2016	FY 2017
Total	68.5	70.9	72.3
Aged	5.5	5.7	5.8
Blind/Disabled	10.5	10.6	10.6
Children	28.0	28.0	28.2
Adults	15.4	15.5	15.7
Expansion Adult	9.1	11.2	12.0
CHIP (avg monthly)³	5.9	6.5	6.7

^a Populations are in millions and may not add due to rounding

^b Preliminary and subject to change

^c Projected estimates

MA - Medicare Advantage
MA PD - Medicare Advantage Prescription Drug Plan
PDP - Prescription Drug Plan
CHIP - Children's Health Insurance Program

downloaded Sep 22, 2017, from [16]

SOURCES: CMS/Office of Enterprise Data & Analytics/Office of the Actuary

MEDICARE COMPONENTS

- Part A or Hospital Insurance (HI) has 98% of persons > 64 years old enrolled in it. Part A services are paid for out of the Medicare Trust Fund
- Part B, or Supplemental Medical Insurance (SMI) covers physician services, and services provided by other types of providers (e.g., health departments); Durable Medical Equipment (DME); and facility charges for hospital outpatient services and ambulatory care centers
- Part C, or Medicare Advantage (HMO, Managed Care) – must have Part A and Part B. Managed Care assumes the risk for beneficiaries
- Part D, or Prescription Drug Coverage was implemented in 2006 as part of the Medicare Modernization Act (MMA) of 2003

Table 2. CMS Financial Data Fiscal Year 2016^d

Total Federal Program Spending (\$ in billions)	\$1,067.2
Medicare Benefits ^e	\$683.6
Total Medicaid	\$368.3
CHIP	\$14.8
Other Spending	\$0.5
Total Program Management (\$ in millions)	\$6,387.7
Total Appropriation	\$4,279.9
Other Sources	\$2,107.8
Total Health Care Fraud & Abuse Control Funding (\$ in millions)	\$1,959.9
FTE Employment	6,238

^d Program Management figures in FY 2016 are preliminary and subject to change.

^e Medicare Benefits, including Health Information Technology Incentive Payments.

downloaded Sep 22, 2017, from [16]

SOURCE: CMS/Office of Financial Management

CMS DATABASE FEATURES

Currently, CMS hosts data in excess of 100 TB with information about approximately 57 million beneficiaries in Medicare and even larger population in Medicaid [16 – 18]. The annual budget for these beneficiaries runs beyond 1 trillion US dollars and is expected to grow exponentially unless checked with proper systemic improvements. More importantly, more than 90 % elderly population and nearly 100 % of critical care patients are covered by CMS databases.

The complexity in the logistics of services provided and the payment services has resulted in 30,000 + individual tables and an apparent list exceeding 8000 variables. Some of these tables have billion + rows and 100s of columns, making their processing quite a challenge. The large number of tables are a result of the efficient partitioning of the data into monthly and yearly categories.

The data itself is made available through different platforms, two major ones of which are: the IDR (INTEGRATED DATA REPOSITORY) and the CCW (Chronic Conditions Data Warehouse). These represent the Database created by CMS integrating, Fee For Service (FFS) claims, Part-D Prescription Drug Events (PDE), Enrollment Database (EDB), Provider of Service (POS) data and Medicare Advantage and Prescription Drug System (MARx). Table 3 shows the number of service providers available in the database.

Table 3. Service Providers in CMS Database ¹⁶	
Provider Type	Number
Total Hospitals	6,146
Total Non-institutional Providers	1,249,691
Total DMEPOS Providers	85,297

CMS has attempted to curb frauds through the Integrated Data Repository (IDR) and One Program Integrity (One PI) programs. While the IDR offers the latest data in the system, CCW alternately referred as CMS Virtual Research Data Center (VRDC) is a more stable version of the integrated data.

Data Linkage is based on several sources like: Census, Cancer registries (e.g., SEER/Medicare), Other providers (e.g., VA, Medicaid), National death index/State vital statistics, Minimum Data Set (MDS), Surveys (e.g. Health and Retirement Study), Provider Information. This linking can take place at the group level based on geography, place of service, etc. At the person level with external data sources like - SSN (plus date of birth and gender) or Medicare ID. A key feature of the CMS database is the unique identifiers for beneficiaries. CMS maintains this unique BENE_ID over time so that the researchers could follow the health status with time:

The IDs used by CMS for Medicare Users have not changed for a long time. Recently there is an attempt to avoid social security number and other vital information to be hidden from the Medicare card. However, the IDs researchers receive have changed over time. HIC (Health Insurance Claim Number) is recently been replaced by BENE_ID. Bene_ID is not SSN or HIC and hence helps to protect the vital information and privacy of the beneficiaries.

ACCESSING CMS DATA

CMS has made its data available at 3 different levels.

1. Public Use File (PUF): These contain non-identifiable DATA and is freely available. It is expected to encourage data entrepreneurs to develop and create software and applications that may eventually be applied to actual CMS claims data; train researchers on the use and complexity of conducting analyses with CMS claims data prior to initiating the process to obtain access to actual CMS data and

support safe data mining innovations that may reveal unanticipated knowledge gains while preserving beneficiary privacy.

2. **Limited Data Set (LDS)** files: are meant for researchers associated with smaller organization with clearly defined DATA needs. They contain beneficiary level information and need DUA (data user agreement). They also have limited variables in comparison with the RIF.
3. **Research Identifiable Files (RIF)**. These are the most detailed data available about the whole population. DUA is a must and often is detailed about covered files, cohort, and years.

MINING AND MANIPULATING DATA

There is structure to this ocean of CMS data. Tables are RDBMS (Relational Database Management System) type, that is, the tables have primary keys to identify the rows uniquely. These keys together make a unique set that link different tables. This RDBMS type structure lends the CMS data to be mined and manipulated by the traditional tools of statistical and database programmers and data scientists. This lends the CMS data for modeling claims, treatments, interventions, innovations etc. The FFS (fee-for-service) claims database, consists of the following seven type of tables:

1. **IP (Inpatient)** files contain final action fee-for-service claims data submitted by inpatient hospital providers for reimbursement of facility charges. All together there are 7 tables listed below with the data divided between them.
2. **The Home Health Agency (HHA)** file contains final action, fee-for-service claims submitted by HHA providers.
3. **Hospice** file contains final action claims submitted by Hospice providers. Once a beneficiary elects Hospice, all Hospice related claims will be found in this file, regardless if the beneficiary is in Medicare fee-for-service or in a Medicare managed care plan.
4. **Outpatient** file contains final action, fee-for-service claims data submitted by institutional outpatient providers. Examples of institutional outpatient providers include hospital outpatient departments, rural health clinics, renal dialysis facilities, outpatient rehabilitation facilities, comprehensive outpatient rehabilitation facilities, and community mental health centers.
5. **The Skilled Nursing Facility (SNF)** file contains final action, fee-for-service, claims data submitted by SNF providers.
6. The **Carrier** file (also known as the Physician/Supplier Part B claims file) contains final action fee-for-service claims submitted on a CMS-1500 claim form. Most of the claims are from non-institutional providers, such as physicians, physician assistants, clinical social workers, nurse practitioners. Claims for other providers, such as free-standing facilities are also found in the Carrier file. Examples include independent clinical laboratories, ambulance providers, and free-standing ambulatory surgical centers.
7. The **Durable Medical Equipment (DME)** file contains final action, fee-for-service claims submitted by Durable Medical Equipment suppliers.

UNDERSTANDING VARIABLES

The complexity of the transactions has necessitated several thousand variables across the whole CMS database. Invariably, one will find similar sounding variables which should be differentiated before selecting the right ones for research. Most confounding ones are probably the date types, corresponding to the various transactions like admission, start, end and payment. In Table 4, we have shown the

number of variables belonging to the seven type of tables. All together one has 1208 variables to work with. Table 4 shows the number variables in the seven class of FFS tables.

Table 4. Number of variables in FFS tables¹⁶	
RIF Files	Number of Variables
Inpatient	271
Skilled Nursing Facility	203
HOSPICE	133
Outpatient	211
Home Health Agency	139
CARRIER	135
Durable Medical Equipment (DME)	116
TOTAL	1208

Following six are special variables that are common to more than one table in the database and come in handy to link the different tables.

1. **Bene_ID:** This is the unique CCW identifier for a beneficiary. The CCW assigns a unique beneficiary identification number to each individual who receives Medicare and/or Medicaid, and uses that number to identify an individual's records in all CCW data. This number does not change during a beneficiary's lifetime and each number is used only once.
2. **CLM_ID:** (Claim ID) This is the unique identification number for the claim. Each Part A or institutional Part B claim has at least one revenue center record. Each non-institutional Part B claim has at least one claim line. All revenue center records or claim lines on a given claim have the same CLM_ID. It is used to link the revenue lines together and/or to the base claim.
3. **CLM_THRU_DT:** The last day on the billing statement covering services rendered to the beneficiary (a.k.a 'Statement Covers Thru Date'). For Home Health prospective payment system (PPS) claims, the 'from' date and the 'thru' date on the RAP (Request for Anticipated Payment) initial claim match. The "thru" date on the claim may not always represent the last date of services, particularly for Home Health or Hospice care. To obtain the date corresponding with the cessation of services (or discharge date) use the discharge date from the claim (variable called NCH_BENE_DSCHRG_DT; note - this variable is not available for Home Health claims). For Part B non-institutional (Carrier and DME) services, this variable corresponds with the latest of any of the line-item level dates (i.e., in the Line File, it is the last CLM_THRU_DT for any line on the claim). It is almost always the same as the CLM_FROM_DT; exception is for DME claims - where some services are billed in advance.
4. **NCH_CLM_TYPE_CD:** Refers to the type of claim that was submitted. There are different claim types for each major category of health care provider.
5. **RNDRNG_PHYSN_NPI** (Rendering Physician NPI): This variable is the National Provider Identifier (NPI) for the physician who rendered the services. This field appears on both the revenue center and base claim files. CMS has determined that dual provider identifiers (old legacy numbers and new NPI) must be available in the NCH. After the 5/07 NPI implementation,

the standard system maintainers will add the legacy number to the claim when it is adjudicated. We will continue to receive the OSCAR provider number and any currently issued UPINs. Effective May 2007, no new UPINs (legacy numbers) will be generated for new physicians (Part B and outpatient claims), so there will only be NPIs sent in to the NCH for those physicians.

6. **RNDRNG_PHYSN_SPCLTY_CD** (Claim or Revenue Center Rendering Physician Specialty Code): The code used to identify the CMS specialty code of the rendering physician/practitioner.

Each Type Of Claim has Multiple Tables Linked

Details of first five of the claims, also referred to as **institutional** claims, are stored in the following seven tables of the respective kinds

1. Base Claim File

This ordinarily contains most details about a particular claim. In case further details are necessary about a claim, the remaining 6 tables can be queried with proper linkage to the base file.

2. Condition Code File

This table has code that indicates a condition relating to an institutional claim that may affect payer processing. Variable CLM_RLT_COND_CD along with other common variables will provide the linkage needed

3. Demonstrations/Innovations Code File

This table contains the details of any CMS demonstration project, the record is part of and denotes any special processing needed (a.k.a. Special Processing Number, SPN). DEMO_ID_NUM, alongside other common variables links the records to individual beneficiaries.

4. Occurrence Code File

This contains the details and code that identifies a significant event relating to an institutional claim that may affect payer processing. These codes are claim-related occurrences that are related to a specific date. The variable CLM_RLT_OCRNC_CD along with the common variables helps in linking with other tables.

5. Revenue Center File

This table contains information about the provider-assigned revenue code for each cost center for which a separate charge is billed (type of accommodation or ancillary). A cost center is a division or unit within a hospital (e.g., radiology, emergency room, pathology). EXCEPTION: Revenue center code 0001 represents the total of all revenue centers included on the claim. It links to the base table with the help of variable REV_CNTR and other common variables.

6. Span Code File

This file contains the details about the significant event relating to an institutional claim that may affect payer processing. These codes are claim-related occurrences that are related to a time period. The variable CLM_SPAN_CD along with the common variables helps in linking with the base claims.

7. Value Code File

This table has details of the monetary condition which was used by the intermediary to process an institutional claim. Variable CLM_VAL_CD along with other common variables will link this to the base claim table

The Carrier and DME files are referred to as **non-institutional** files and have their data linked across three different tables: Base Claim File, Demonstrations/Innovations Code File and Line File. The first two are the same as described before. The Line File is linked by the variable CLM_LN (claim line number) in conjunction with other common variables.

LIMITATIONS:

Mues et al have reviewed the strength and limitations of the Medicare data [17]. ICD-9 diagnosis code doesn't provide enough detail and since October 2015 CMS has implemented ICD-10 standards. Some of the ICD-9 limitations restricted the use of the data. For e.g.: - Cancer diagnosis can be found as an ICD-9 diagnosis code (e.g., lung cancer is 162.xx) but not stage or histology. No indication of incidence, severity can be inferred. Knowing that someone has a chronic disease (e.g., congestive heart failure) doesn't provide information about how long they have had the condition (incidence vs. prevalence) or severity. Data Quality Variability exists due to the data being typically limited to covered benefits and for which claims are submitted. No information about Part B services for Managed Care enrollees is available. Further, little information (of largely unknown quality) about hospitalizations for managed care enrollees. Not all beneficiaries have Part D coverage; not all beneficiaries with Part D coverage will have Part D utilization information contained in files

In general, as the data is based on the payment system, General Variable Quality is:

- ◆ If it impacts payment its quality will be better than if it doesn't.
- ◆ Different types of care may be subject to different payment rules

This further implies that:

- ◆ Comorbidity and severity of illness information may be inconsistently recorded
- ◆ Some components of treatments may not be included in bills if reimbursement rates are very low (even if the treatment is provided)

CONCLUSION

CMS Data is available through multiple platforms like IDR and CCW. The data is well structured and robust with information about a large section of US population otherwise not available and can be expected provide a vital supporting role to the traditional healthcare research. Unrestricted, Public Use Files are available for exploratory work and detailed RIF data is available for further research. Knowledge of the vast list of variables is important for project success. Data updates are regular and should be considered when comparing historic results. Since 2015, ICD10 codes are implemented and should lead to more opportunities for detailed research.

Useful CMS links:

- <https://www.cms.gov/Research-Statistics-Data-and-Systems/Research-Statistics-Data-and-Systems.html>
- <https://www.resdac.org/>
- <https://www.ccwdata.org/web/guest/data-dictionaries>
- <https://www.resdac.org/cms-data/file-availability>

- <https://www.cms.gov/Medicare/Coding/ICD10/2016-ICD-10-CM-and-GEMs.html>

REFERENCES

1. S Schneeweiss and Jerry Avorn, *Journal of Clinical Epidemiology* 58 (2005) 323–337
2. S. E. Harpe, *The Journal of Human Pharmacology and Drug Therapy*, 29(2009) 138–153. Using Secondary Data Sources for Pharmacoepidemiology and Outcomes Research. doi:10.1592/phco.29.2.138
3. K. Verhamme, M. Sturkenboom, *European Journal of Clinical Pharmacology*, 2010, 67 (S1), pp.67-74.
4. <https://www.healthdata.gov/content/about>.
5. S. V.Wang, S. Schneeweiss, M. L.Berger, J. Brown, F. deVries, I. Douglas, J. J.Gagne, R. Gini, O. Klungel, C.D. Mullins, M. D. Nguyen, J. A. Rassen, L.S. , MiriamSturkenboom, *Value In Health* (2017) In Press. <http://dx.doi.org/10.1016/j.jval.2017.08.3018>
6. E. I. Benchimol, L. Smeeth, A. Guttman, K. Harron, D. Moher, I. Petersen, H. T. Sørensen, E. von Elm , S. M. Langan.*PLoS Med*2015;12:e1001885.
7. S. G. Nicholls, P. Quach, E. von Elm, A. Guttman, D. Moher, I. Petersen, H. T. Sørensen, L. Smeeth, S. M. Langan , and E. I. Benchimol , Published: May 12, 2015. <https://doi.org/10.1371/journal.pone.0125620>
8. Altman D. G, Simera I, Hoey J, Moher D, Schulz K (2008) EQUATOR: reporting guidelines for health research. *Lancet* 31: 1149–1150.
9. D. B. Rubin, *Journal of Educational Psychology*, 66(1974), 688-701. <http://dx.doi.org/10.1037/h0037350>
10. Gi-Soo Kim, M.Cho Paik, H. Kim, *Computational Statistics & Data Analysis*, 113 (2017), 88-99.
11. H. Kim, W. W. Hung, M. Cho Paik, J. S. Ross, Z. Zhao, Gi-Soo Kim, K. Boockvar, *International Journal for Quality in Health Care*, 27 (2015), Issue 6, 513–519. <https://doi.org/10.1093/intqhc/mzv082>.
12. A. Melamed, J. A. Rauh-Hain, J. O. Schorge, *Gynecologic Oncology*, 146(2017) 653–660
13. I.Simera, D. G. Altman, D. Moher, K. F. Schulz, J. Hoey, *PLoS Medicine* 5(2008). doi:10.1371/journal.pmed.0050139.t001
14. Erik von Elm, Douglas G. Altman, Matthias Egger, Stuart J. Pocock, Peter C. Gøtzsche, and J. P. Vandenbroucke, *Epidemiology* 18(6), November 2007, 800-804. DOI: 10.1097/EDE.0b013e3181577654
15. G. Migali, E. Zucchelli, *Journal of Economic Psychology* 62 (2017) 98–119
16. <https://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/CMS-Fast-Facts/index.html>; downloaded on Sep 22, 2017.
17. K. E. Mues, A. Liede, J. Liu, J. B. Wetmore, R. Zaha, B. D. Bradbury, A. J. Collins, D. T. Gilbertson, *Clinical Epidemiology* 9(2017) 267–277.
18. <https://www.resdac.org/cms-data/file-family/Medicare-Claims>
19. <https://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/CMSProgramStatistics>