Health Analytics for Medicaid Claims Data

The Health Analytics group at Georgia Tech conducts research in data science to improve decision making in health care delivery and public health. Georgia Tech (with support from IPaT, ISyE, and Children's Healthcare of Atlanta) purchased Medicaid claims data for multiple states and years to be used for health analytics research. Additional data sets have also been obtained for research.

What kind of Medicaid data do you have?

We have the Medicaid Analytic eXtract (MAX) files that are person-level, Research-Identifiable-Files. The five file types are listed below. A National Provider ID (NPI) and Characteristics file is available for 2009 and later years, which provides additional information on providers. The files include claims paid for patients under managed care organizations and fee-for-service plans. More information is available from CMS. http://www.cms.gov/Research-Statistics-Data-and-Systems/Computer-Data-and-Systems/MedicaidDataSourcesGenInfo/MAXGeneralInformation.html

1. **Personal Summary**: patients, demographics, birthdate, etc.
2. **Inpatient**: claims, diagnoses, procedures, LOS, payment
3. **Other Therapy**: claims for physician, lab, clinic, outpatient
4. **Long Term Care**: facility type, date of service, etc.
5. **Prescription Drug**: paid drug claims

What populations, states, and years are available?

The initial Data Use Agreement and approved research protocol covers all Medicaid beneficiaries who are **children or pregnant women**. The initial data obtained covers **years 2005 – 2009** inclusive for **14 states including those in the southeast (Georgia, Alabama, Arkansas, Louisiana, Mississippi, N. Carolina, S. Carolina, Tennessee, Texas) and comparison areas (California, Minnesota, New York, Pennsylvania)**.

Georgia Tech has initiated the process for purchasing additional data and with expanded populations. This data request covers **all 50 states, years 2010 and 2011, and all beneficiaries under Medicaid**. Additional years will be purchased when they are available. The request is to purchase four file types including Personal Summary, Inpatient, Other Therapy, and Prescription Drug.

How can researchers access the data?

The MAX claims data includes confidential information and is protected by a Data Use Agreement between Georgia Tech and CMS. The existing Data Management Plan only allows direct access to the raw data and non-aggregated data by employees and students of Georgia Tech who have also undergone appropriate training and IRB approval.

Data can be extracted and aggregated on-site at Georgia Tech, and data with a sufficient level of aggregation can be provided to collaborators. At a minimum, the shared data must contain at least 11 entries in each cell and in the numerator and denominator of ratios. Data is also reviewed at GT before release to ensure identification of patients or providers across multiple data sets is not possible.

How can scientists from Children's Healthcare of Atlanta, Emory, or other institutions use the data for research?

Children's or Emory scientists (or those from other institutions) can partner with Georgia Tech on specific research questions. There are different levels of engagement depending on the level of expertise needed. The most basic data extraction involves a joint understanding of specific data elements, rules to join or group by, etc., and it may involve multiple iterations to obtain the right data set. More sophisticated analysis requires using the de-aggregated information available at the patient or provider level and must be done by GT employees in the data room. The GT group has been developing many queries, algorithms, and analysis to understand the data and can work with you to define the best approach.

If you have a specific project you would like to explore, please email phdi@gatech.edu (and/or Sherry Farrugia) with a short description of the research. You will be contacted to explore in further detail.

Does doing research with this data cost money?

The short answer is yes, everything at research universities involves cost, including for data, infrastructure to house it, information technology support, researchers, etc. The specific cost would depend on the type of data needed and the resources involved.

If you have existing grants, then you may be able to use those. You may also be able to apply for funding from programs like those listed at Pediatric Connects https://pediatriconnect.gtri.gatech.edu/grants or others that may be available to you at Children's or Emory. Even if you do not have a funding source right now, you are encouraged to contact the group at phdi@gatech.edu with your project idea in case funds become available.

Can other data be combined with the MAX Medicaid claims?

The existing research protocol defines the types of data that can be connected to the Medicaid claims within the protected data room. For example, there is approval to connect the National Provider ID file, census information, and geographical information. At present we do not have approval to connect individual patients or individual providers in Electronic Health Records with the Medicaid claims data.

After the data is aggregated to a sufficient level and is released from the data room, it can also be connected with other data (e.g., geographical). All researchers should keep in mind that CMS has requested that confidentiality be maintained on both patients and providers.

What research topics can be studied?

The initial protocol approves research with the below aims. The second data reuse expands the same questions to the overall population on Medicaid.

1) MEASURING AND EXPLAINING INEQUITIES:
    a. To assess the impact of healthcare system characteristics vs. inequities in healthcare, including geographical, use, quality, expenditure and outcomes among Medicaid children enrollees, especially in states with historic inequities like in the southeast.
    b. To identify geographic areas with widespread and increasing Medicaid healthcare use (status quo and over time) and determine the underlying associative factors (e.g. access to healthcare facilities, race and ethnicity);
    c. To investigate geographic variations in healthcare quality indicators (adherence to medications, emergency room visits, and other utilization measures) for high-impact

diseases in children such as respiratory deficiencies, obesity, diabetes and other disabilities;

d. To identify geographic subdivisions which have achieved good health outcomes and low disparities despite adverse social determinants, or which have achieved poor health outcomes and high disparities worse than the social inequalities.

2) OPTIMIZING INTERVENTIONS AND DELIVERY SYSTEMS

a. To analyze flows and policies across the system, e.g., the match between supply and demand, and financially, both geographically and across time, along with the corresponding costs or outcomes, to analyze improved methods of delivery including medical homes.

b. To examine areas in the children Medicaid expenditure with the greatest costs or utilization, and assess potential interventions for reducing the healthcare costs, especially where interventions may be targeted by patient characteristics such as risk or where chronic issues like pediatric obesity can be addressed;

c. To evaluate the potential costs and benefits to creating a medical home or using telemedicine in the Medicaid system, where the creation may be focused on a subgroup of the Medicaid population or within specific geographical areas with great need;

d. To forecast the available "supply" of general or specialist providers or network services across geographical regions (e.g., counties or census tracts) as a function of socio-economic and other elements, link this factor with the costs or outcomes in the system as measured by the claims data, and examine potential interventions.;

e. To evaluate the impact of various public policies, such as changes in cost-sharing, on the demand for Medicaid coverage.

Can other research questions be studied?

Topics outside of the scope of approved protocol require researchers to submit a request to the Center for Medicare and Medicaid Services for Data Reuse and go through a review by an Institutional Review Board and a CMS entity. Researchers are encouraged to identify research questions aligned with the aims described above.

Does the GT Health Analytics group use other data for research?

Absolutely! We use data of many types and sources including Electronic Health Records, registry databases for a specific disease, national or state surveys like NHANES and BRFSS, sets from the Hospital Cost and Utilization Project, Clinical Risk Grouping provided by 3M software, infectious disease surveillance monitoring, geographical information, and many others. Some examples are included on our webpage. http://www.healthanalytics.gatech.edu/ Great value can be achieved by combining information gleaned from multiple sources.

I am working with someone else at Georgia Tech; do they have access to the data?

Existing GT researchers using the data include Nicoleta Serban and Julie Swann. There are others currently approved on the protocol (e.g., Rahul Basole, Dave Goldsman, Pinar Keskinocak) who can use the data. Other GT faculty and scientists can request access to the data through a process described online (INSERT RICHARD's link here).

What else do I need to know?

The MAX claims data is a rich data set that allows the study of many research questions. There are some limitations inherent with using claims data, and the data from some states or data elements is better than others. In general the GT group has found that lots of interesting and useful information can be generated from the data sets. In addition, the GT group is eager to collaborate with scientists who have specific research questions that the data can help answer.