**Challenges of Analytics in Healthcare**

The healthcare sector is a prime candidate for use of data analytics because of the complexity and social impact of the field. Health informatics suffers characteristic challenges of big data[2] due to the volume, velocity, variety, and veracity of the associated data. Typical analytics methodologies can be applied to healthcare data via problem proposal, data collection/ETL, analysis, results and insight, and deployment. Although our project does not necessarily involve big data sources, we will be taking the same approach to provide a framework for visualizing healthcare data that others can utilize.

The Open Government Strategy[10] is aimed at improving access to healthcare data. As of 2010, the Centers of Medicaid & Medicare Services (CMS) have been providing publically available data, which researchers can use to inform the public and improve health outcomes. Since Medicaid data is not readily interpretable, our project aims to visualize data for pediatric diseases on a publically available site so that others can easily access it.

Outlined below are some use cases for analytics in healthcare. We will be implementing similar techniques and methodologies for our project.

**Use Case 1: Patient Management and Profiling, Financial Analysis**

Data analytics techniques and tools are indispensable for managing patient records, diagnoses and treatments, and claims. Patient profiles[3] can be developed via segmentation and predictive modeling to identify individuals who would benefit from proactive care to prevent adverse outcomes. Analytics can also be used to shed light on financial issues[11] in the healthcare field, such as by flagging missing charges, scheduling, and determining which parts of the healthcare flow are causing the most financial strain on the system.

**Use Case 2: Public Health, Epidemiology**

Analytics can be used in public health to track, model, and predict disease prevalence, outbreaks/transmissions, and outcomes so as to improve public health surveillance and response speed[2]. HealthMap[8] is a tool that monitors global infectious disease through automated classification and visualization of internet media reports. We can apply similar algorithms to detect trends and change within our data set. Spatial epidemiology[9] is a subdivision of the field that focuses on disease mapping, disease clustering, and geographical correlation analysis via dot maps for case-events, choropleth maps for regional data, and isopleth maps for geostatistical data.

**Topic of Focus: Pediatric Chronic Conditions**

Due to population growth and overall average age increase, there have been significant trends[4] toward chronic, non-communicable diseases. Disease trends provide an important geographical indication of whether a society is making progress in advancing positive health outcomes. The Clinical Risk Group system is used to classify children into health status groups according to Episode Diagnostic Category classification[1]. A study on emergency department visits of children[12] looks at how the number of chronic conditions affects total yearly cost of the emergency department, and which diseases when paired together cost the most to treat. This is useful for our project because it uses approach similar to what we plan on doing to analyze overall cost. We will try to improve on this idea by using a regression model to forecast expected cost based on the geographical location and chronic condition of the patient.

**Method: Data Visualizations**

Using data visualization techniques instead of database techniques for healthcare data is often preferred[7] because visuals can efficiently communicate complex ideas and they allow for interactive presentation when implemented properly. Modern tools enable researchers to easily publish and disseminate data to the public via simple web platforms[8]. For large complex datasets, it is difficult to have a one-size-fits-all visualization. An example web-based visualization tool is CoeViz[5], which combines several different visualizations to provide coevolution analysis of protein residues. While our datasets are not as complicated as amino acids, we can use similar techniques to build interactive heatmaps, charts, and other visuals to communicate the data. Other popular and prolific web visualization tools include D3[6], HighCharts, and IVIS which we will utilize to build dynamic, interactive data visuals.

**Project Proposal**

This project aims to analyze census data and create reusable visualizations to display disease prevalence and associated healthcare costs. The main objective is to visualize trends in prevalence and costs of 25 different pediatric diseases stratified by geography and time. The final deliverable will be a choropleth of prevalence metrics for several chronic diseases in children along with a prediction of incurred cost to the Medicaid system.

Currently, the CDC has data and metrics on major chronic disease indicators (e.g. cancer, heart disease, diabetes, and stroke), but there is significantly less data and exposure on the pediatric diseases that are the focus of this project.

Our new approach is to link the pediatric prevalence data with associated costs on a single map. Then using the cost and prevalence data, we will forecast expected cost in the upcoming year using a regression model. This project will be successful because it will provide exposure to data that is not currently visualized elsewhere.

Stakeholders who would be interested in this project include healthcare professionals and groups, such as GT Health Analytics Group, Children’s Healthcare of Atlanta (CHOA), Center for Disease Control (CDC).

If successful, this project will publish easily accessible and understandable visualizations for the GT healthcare analytics group. We can measure success of the visualizations by looking at site usage statistics as well as number of references. The payoff for this project is a unique tool that GT researchers can utilize to easily visualize their data and results, thereby gaining more exposure to the subject matter and potentially receiving more funding. Some risks for this project include poor regression model forecasts and difficulty obtaining or working with the data.

Since we plan on using open source technologies and public datasets created by GT research groups, there will be no incurred costs other than time.

This project will take approximately 7 weeks for activities including research, data ETL, visualization, and final presentation. At the mid-project checkpoint, data cleaning, modeling, and analysis should be complete. The final finished product will be all visuals and analysis published to the GT healthcare analytics website.

**Activity Plan**

|  |  |  |  |
| --- | --- | --- | --- |
| **Team Member** | **Completed** | **Assigned Tasks** | **Est. Time** |
| Alex | Data gathering | * Get more data (cost, previous years) * Data cleaning | 2 weeks |
| David | See all | * Research on visualization methodologies * Regression analysis | 2 weeks |
| Sarah | Project proposal, project management | * Research medical data * Data modeling | 2 weeks |
| Tim | Video | * Data analysis * Build visualizations | 3 weeks |
| All | Literature review  Proposal ppt | * Build visualizations * Progress report, final report | 3 weeks |

**Distribution of team member effort**:

**References**

[1] Neff, J., Sharp, V., Muldoon, J., Graham, J., Popalisky, J., Gay, J. (2002).“Identifying and Classifying Children With Chronic Conditions Using Administrative Data With the Clinical Risk Group Classification System.” *Ambulatory Pediatrics*, *2*:71-79. Retrieved from <http://www.dhcs.ca.gov/services/ccs/Documents/IDCronicConditions.pdf>

[2] Raghupathi, W.; Raghupathi, V. (2014). "Big Data Analytics In Healthcare: Promise And Potential." *Health Information Science and Systems*, *2*:3-13. Retrieved from <https://www.biomedcentral.com/track/pdf/10.1186/2047-2501-2-3?site=hissjournal.biomedcentral.com>

[3] [Bates, DW](https://www.ncbi.nlm.nih.gov/pubmed/?term=Bates%20DW%5BAuthor%5D&cauthor=true&cauthor_uid=25006137), [Saria, S](https://www.ncbi.nlm.nih.gov/pubmed/?term=Saria%20S%5BAuthor%5D&cauthor=true&cauthor_uid=25006137)., [Ohno-Machado, L](https://www.ncbi.nlm.nih.gov/pubmed/?term=Ohno-Machado%20L%5BAuthor%5D&cauthor=true&cauthor_uid=25006137)., [Shah, A](https://www.ncbi.nlm.nih.gov/pubmed/?term=Shah%20A%5BAuthor%5D&cauthor=true&cauthor_uid=25006137), [Escobar, G](https://www.ncbi.nlm.nih.gov/pubmed/?term=Escobar%20G%5BAuthor%5D&cauthor=true&cauthor_uid=25006137). “Big Data In Health Care: Using Analytics To Identify And Manage High-risk And High-cost Patients." *Health Affairs, 7*:1123-31. Retrieved from <https://www.ncbi.nlm.nih.gov/pubmed/25006137>

[4]

Lozoano, R.; Naghavi, M.; Foreman, K; Lim, S.; Shibuya, K.; Aboyans, V… Murray, C. (2012) “Global And Regional Mortality From 235 Causes Of Death For 20 Age Groups In 1990 And 2010: A Systematic Analysis For The Global Burden Of Disease Study 2010.” *The Lancet, 380*:2095-2128. Retrieved from <http://dro.deakin.edu.au/eserv/DU:30050819/ortblad-globalandregional-2012.pdf>

[5] Baker, F.; Porollo, A. (2018) “CoeViz: A Web-Based Integrative Platform for Interactive Visualization of Large Similarity and Distance Matrices” *Data*, *3*(1): 4. Retrieved from <http://www.mdpi.com/2306-5729/3/1/4/htm>

[6] Lu, Z.; Zhang, Y, (2017) “Facilitated Analysis Of Large Data Sets By Interactive Visualisation.” *BioRxIv*. Retrieved from <https://www.biorxiv.org/content/biorxiv/early/2017/08/20/178616.full.pdf>

[7] Carroll, L. N., Au, A. P., Detwiler, L. T., Fu, T., Painter, I. S., Abernethy, N. F. (2014). “Visualization And Analytics Tools For Infectious Disease Epidemiology: A Systematic Review”. *Journal of Biomedical Informatics*, 287-298. Retrieved from<https://www.sciencedirect.com/science/article/pii/S1532046414000914#f0030>

[8] Freifeld, C. C., Mandl, K. D., Reis, B. Y., & Brownstein, J. S. (2008). “HealthMap: Global Infectious Disease Monitoring through Automated Classification and Visualization of Internet Media Reports.” *ScienceDirect*, 15(2): 150-157. Retrieved from https://academic.oup.com/jamia/article/15/2/150/707275

[9] O. B. (2003). “Exploratory Disease Mapping: Kriging The Spatial Risk Function From Regional Count Data.” *International Journal of Health Geographics, 3*:18. Retrieved from<https://ij-healthgeographics.biomedcentral.com/articles/10.1186/1476-072X-3-18>.

[10] Conway, P., & VanLare, J. (2010). “*Improving Access to Health Care Data*.” *The Journal of the American Medical Association*, 304(9): 1007.

[11] Simpao, A., Ahumada, L., Gálvez, J., & Rehman, M. (2014). “A Review of Analytics and Clinical Informatics in Health Care.” *Journal of Medical Systems*, 38(4): 45.

[12] Berry, J., Rodean, J., Hall, M., Alpern, E., Aronson, P., Freedman, S., . . . Neuman, M.. (2017) “Impact of Chronic Conditions on Emergency Department Visits of Children Using Medicaid.” *The Journal of Pediatrics*,182: 267-274.