Clinical Pearl: "What Does Research Analysis of "Big Data" Really Mean to the Pediatric Provider Regarding Sudden Unexpected Infant Death and Safe Sleep Recommendations?

Tatiana M. Anderson, PhD and Kelty Allen, PhD

The digitization of population health data has led to a large amount of "big data" and stimulated a rapid growth of data science in medicine. Data is generated in high volume, in a large variety of formats, and amasses so fast that it cannot be analyzed using traditional data-processing methods (1). The combination of big data, the exponential growth of computational power (2) and the availability of open-source programming tools together with advances in machine learning and analytics techniques have created the perfect storm for the growing importance of big data in healthcare. Turning this firehose of data into meaningful insights can lead to better decision making in the clinic, provide higher quality patient care, and ultimately save lives.

The literature contains many papers on the value and vision of big data in medicine (3-7) and is rife with real-world applications exemplifying the power of data science especially in the fields of medical image processing, signal processing, and genomics (8,9). Here, we do not wish to add another overview of the benefits of big data in healthcare and instead offer the reader a non-technical, step-by-step understanding of the processes by which our team uses large datasets to lead to actionable insights for medical professionals in an effort to decrease instances of sudden infant death syndrome (SIDS) and sudden unexpected infant death (SUID). This has been possible through a unique collaborative partnership between researchers at Seattle Children's Research Institute, data scientists at Microsoft, and top medical researchers in the SIDS/SUID field.

"This has been possible through a unique collaborative partnership between researchers at Seattle Children's Research Institute, data scientists at Microsoft, and top medical researchers in the SIDS/SUID field."

Our primary data source has been the Birth Cohort Linked Birth – Infant Death data files publicly available through the Centers for Disease Control's National Center for Health Statistics (10). A data scientist starts by uploading data from every live birth in the United States for any number of years between 1983-2014 (the most current year available). This includes a total of over 10 million live births per year. If there is an early death, the death certificate is linked to the birth certificate, so we can search for cause of death codes established in the International Classification of Diseases 10th edition (ICD-10) that are specific to the umbrella category 'SUID' including SIDS (R95), unknown or ill-defined cause of death (R99), and accidental strangulation or suffocation in bed (W75). There are approximately 3,500 cases of SUID in the United States annually. While instances of a SIDS diagnosis have decreased, the number of R99 and W75 diagnoses have increased (a phenomenon called 'diagnostic shift' or 'diagnostic transfer' (11)) such that overall SUID cases have been stagnant for the last couple of decades.

After uploading, the data is stored in a cloud-based database using Microsoft Azure. The publicly available data is deidentified, meaning that it does not include any private, personally identifying information such as names, addresses, geographic locations, etc. Data scientists can then pull subsets of the data from the database and use programming languages, such as Python or R, and open-source libraries to begin preliminary analysis.

Whether a specific hypothesis is in mind or not, the first step involves data exploration to start to learn patterns and look for anomalies or missing data. At this stage the data scientist is often in close communication with colleagues and medical experts to better formalize the hypothesis, question, and best analysis method for obtaining results. The goal is to produce novel findings that are, of course, clinically relevant.

Some special techniques are necessary due to the fact that SUID is a rare event. In 2017 0.09% of live births died of SUID in the US. If the data scientist wanted to build a model that, for example, predicted SUID risk by the number of cigarettes a mother smoked during pregnancy, the model could not be trained using the whole population. The machine learning model optimizes for accuracy, precision, and recall for predicting 'SUID' vs. 'not SUID', and thus the best model would "learn" to predict 'not SUID' on every data point giving an accuracy rating of 99.91%! To avoid this issue, we use a technique called down-sampling, wherein the model uses a representative sample of non-SUID infants about 10 times the size of the SUID population. This type of model cannot be used to "predict" SUID in the real world, however, statistics such as adjusted odds ratio are reliable.

Writing code can be tedious, but running the code is often very fast. Because of this, it is usually worth it to try several kinds of models to see what works best with the data. For example, if we are trying to understand which features are associated with increased or decreased risk of SUID or calculate the adjusted odds ratio of SUID in one population compared to another, our team has used logistic regression, generalized additive models, and Bayesian networks as well as other standard classifiers like random forest and support vector machines.

From here, we work to compile a valuable story, often consulting with pediatricians, epidemiologists, and medical researchers to get feedback on what is interesting, novel, and actionable. In other words- how do we go from "statistically significant" to "this is what this means for your patients".

One of the best examples of this journey end-to-end is a manuscript that was published by members of our team in April 201912.

The association between maternal smoking and an increased risk of SUID has been well-documented, however, since we are working with such large data sets, we were able to define this relationship with much higher, single-cigarette resolution. We found that smoking a single cigarette a day, on average, during pregnancy doubles the risk of SUID. There was a linear increase in risk between smoking 1 and 20 cigarettes. Smoking in the 3 months before pregnancy and quitting by the first trimester still increased SUID risk by approximately 50%. Decreasing cigarette consumption during pregnancy decreased SUID risk. Through this story we were able to provide solid, actionable data and asked clinicians to 1) urge patients to quit smoking well before trying to become pregnant, 2) unequivocally tell patients that the best way to lower SUID risk is to quit smoking, and 3) tell patients that refuse or are unable to quit that even reducing the number of cigarettes makes a positive difference.

Thanks to this incredible partnership, we currently have two published manuscripts (12,13), two in the review process, and several in various stages of analysis/drafting. In an effort to continue the momentum and foster new ideas and collaborations, Seattle Children's Hospital and Microsoft jointly host an annual "SIDS Summit", a two-day conference dedicated to SIDS, SUID, and infant mortality in Seattle. We bring together researchers, data scientists, clinicians, geneticists, pathologists, medical examiners, and bioinformaticians to present their latest research and tackle the issues from multiple angles of expertise.

We believe that this partnership can serve as a template model for a host of other issues, for example diabetes, cancer, or the current opioid epidemic. With ever-increasing amounts of medical data being collected and data science as one of the fastest growing career fields, the time is ripe to leverage big-data to help drive evidence-based care and prevention.

References

1. *Sanchez-Pinto LN, Luo Y, Churpek MM. Big data and data science in critical care. Chest. 2018;154(5):1239-1248.*
2. *Lundstrom M. Moore's law forever? Science. 2003;299(5604):210-211.*
3. *Obermeyer Z, Emanuel EJ. Predicting the future—big data, machine learning, and clinical medicine. The New England journal of medicine. 2016;375(13):1216.*
4. *Raghupathi W, Raghupathi V. Big data analytics in healthcare: promise and potential. Health information science and systems. 2014;2(1):3.*
5. *Chaussabel D, Pulendran B. A vision and a prescription for big data–enabled medicine. Nature immunology. 2015;16(5):435.*
6. *Alyass A, Turcotte M, Meyre D. From big data analysis to personalized medicine for all: challenges and opportunities. BMC medical genomics. 2015;8(1):33.*
7. *Austin C, Kusumoto F. The application of Big Data in medicine: current implications and future directions. Journal of Interventional Cardiac Electrophysiology. 2016;47(1):51-59.*
8. *Belle A, Thiagarajan R, Soroushmehr S, Navidi F, Beard DA, Najarian K. Big data analytics in healthcare. BioMed research international. 2015;2015.*
9. *Costa FF. Big data in biomedicine. Drug discovery today. 2014;19(4):433-440.*
10. *Centers for Disease Control and Prevention National Center for Health Statistics. Vital statistics data available online. Cohort linked birth-infant death data files. Available at:* www.cdc.gov/nchs/data_access/Vitalstatsonline.htm.
11. *Byard RW. Changing infant death rates: diagnostic shift, success story, or both? : Springer; 2013.*
12. *Anderson TM, Lavista Ferres JM, Ren SY, Moon RY, Goldstein RD, Ramirez JM, et al. Maternal Smoking Before and During Pregnancy and the Risk of Sudden Unexpected Infant Death. Pediatrics. 2019;143(4).*
13. *Lavista Ferres JM, Anderson TM, Johnston R, Ramirez JM, Mitchell EA. Distinct Populations of Sudden Unexpected Infant Death Based on Age. Pediatrics. 2020;145(1):e20191637.*

Conflicts: The authors have no conflicts to disclose

NT

Corresponding Author

Tatiana M. Anderson, PhD
Neuroscience Postdoctoral Fellow
Seattle Children's Research Institute
Center for Integrative Brain Research
Tatiana A <tatiana.m.anderson@gmail.com>

Kelty Allen, PhD
Data Scientist at Stripe
San Francisco, CA