To Your Health!

Joe Kincaid
Blue Cross and Blue Shield of Kansas City
Business Process Engineer, Senior

In 2001, researchers at the Framingham Heart Study published results [1] describing how diastolic blood pressure, systolic blood pressure, and pulse pressure can be used as predictors of coronary heart disease risk. This study was able to improve our ability to predict coronary heart disease by analyzing the strength of the predictors for different age groups, leading to a conclusion regarding which predictor was strongest for each age group.

Predictors for health and wellness are improving in many areas for many reasons. The available data is expanding dramatically. The technology and techniques we use for modeling and analysis are improving. Advances in one of these areas will in turn contribute to advances in the other areas. All of this enables researchers to reach deeper, more insightful conclusions than could be reached previously.

Data

The Framingham Heart Study began in 1948 and has expanded several times since then adding cohorts reflecting not only subsequent generations, but also the increased diversity of the population. [2] This allows the researchers to use data that represents both the breadth of the population as well as the longitudinal aspect of many health questions. For the study described above, the researchers combined the original cohort with the offspring cohort providing an age distribution from 20 years to 79 years.

Researchers are also collecting data in places that would not have been possible a few years ago. Fitness trackers easily quantify information about a person's life, such as activity level and sleep patterns. Collecting this data would have required more intrusive methods just a few years ago. Researchers are using fitness data to better understand how fitness impacts our risk for various medical conditions. For example, researchers are using fitness as a predictor for the risk for stroke [3] and time spent being sedentary as a predictor for type 2 diabetes [4]. Consumers with better fitness levels have been shown to have better chances for surviving a first heart attack [5].

Technology

Technology improvements are facilitating these advances in many ways. Computers continue to increase in storage capacity, processing capabilities, and data throughput. The overall cost of the same level of computing is decreasing with each round of advances. According to the NSF, the percentage of academic institutions with 10Gbps or greater network connections has increased from 2% to 25% over the last 10 years [6]. Over 35% of the institutions in that 2011 study reported high-performance computing resources of 1 teraflop or faster.

Combining new data with new technologies provides opportunities to use big data analysis. For example, clinical data, genomic data, and environmental data are being combined so that interactions

between them can be studied [7]. This analysis would not have been possible without the computing capacity to manage the large data sets found in bioinformatics.

Modeling

Having more data and better technology also allows us to expand or repurpose previous models in new ways. For example, Bijak and Bryant recently published a paper [8] in *Population Studies* showing how Bayesian methods can be used to reach new conclusions in demographic analysis. In that paper, they speak to recent improvements in computing the posterior distribution. There is still some work to be done in this area, but software packages are improving.

Bijak and Bryant reference work by Gelman and Hill in describing how Bayesian analysis is well suited for situations where data are sparse or incomplete. By selecting a hierarchical model where the likelihood is based on the estimate from the data and the prior distribution is based on a model for the underlying situation, the posterior distribution will give more weight to the likelihood in cells where the data has more observations and more weight to the prior in cells where the data has fewer observations. This leads to a compromise where "the hierarchical model is sensitive where it can be, and robust where it needs to be" [9].

Analysis

Not surprisingly, as theoretical methods improve, we see advances in our ability to make predictions as well. The 2001 study using the Framingham Heart Study population cited above used Cox regression to estimate the relation between the blood pressure indices and the risk of coronary heart disease. This technique was first published in 1972 by D. R. Cox [10].

More recently theoretical advances allow geographic information systems to be combined with data on outbreaks to better model epidemics and predict the transmission of infectious diseases [11]. The bioinformatics study cited in [7] is using Bayesian networks, a modern statistical method suitable for analyzing big data sets from disparate sources.

By combining all of this, the medical community can better understand prospects for one's future health based on current health, current lifestyle, and other indicators such as family history. This is a team effort relying on data scientists, medical researchers, computing infrastructure, and theoretical work by mathematicians and statisticians. Each of these areas benefits from advances in the others. As access to data increases, for example, greater computing capacity for calculating with that data is needed. This is an exciting time for health care research.

References

[1] Framingham Heart Study http://circ.ahajournals.org/content/103/9/1245.full
[2] https://www.framinghamheartstudy.org/about-fhs/index.php
[3] http://consumer.healthday.com/sleep-disorder-information-33/misc-sleep-problems-news-626/getting-enough-sleep-exercise-may-lower-stroke-risk-708117.html
[4] http://www.medicaldaily.com/type-2-diabetes-risk-every-extra-hour-sitting-371872
[5] http://consumer.healthday.com/fitness-information-14/misc-health-news-265/fitness-likely-to-boost-survival-rate-after-first-heart-attack-707476.html
[6] http://www.nsf.gov/statistics/infbrief/nsf13329/
[7] http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4076480/
[8] http://www.tandfonline.com/doi/full/10.1080/00324728.2015.1122826
[9] Gelman, Andrew and Jennifer Hill. 2006. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge: Cambridge University Press.
[10] http://www.jstor.org/stable/2985181?seq=1#page_scan_tab_contents
[11] http://www.cell.com/trends/parasitology/abstract/S1471-4922(16)00002-7