

Predictive Modeling NEWS

Socioeconomic Data Fuels Deeper Understanding of Health Outcomes, Stress Levels, Motivation – Especially for New Enrollees

Many predictive analytics organizations are recognizing the value of integrating non-clinical data – primarily socioeconomic data from outside sources – into the analytics that feed a patients' care plan. LexisNexis has been a leader in that type of data integration and in applying it to population health management and other predictive modeling-driven endeavors. Predictive Modeling News talked to Ogi Asparouhov PhD, chief data scientist there, about the value of incorporating socioeconomic data in predictive analytics.

Predictive Modeling News: What was the spark that led to incorporating socioeconomic data in existing stores of clinical and claims data? Was it the obvious next step in the development and use of "Big Data"? Was there a disruptive event that made it happen?

Ogi Asparouhov PhD: The Affordable Care Act really boosted the development and use of Big Data and analytics across healthcare; but our strategy has always been to find innovative ways to improve how healthcare is administered and delivered. The rush of consumers entering the health system who are low or non-utilizers makes finding and using new data sources to assess and predict risk that much more important.

The healthcare outcome predictive modeling industry relies on a good mixture of four components: clinical, healthcare, mathematical and computer knowledge. In my opinion, the HOPM industry's first 25 years – 1980 to 2005, which I call "the risk adjustment methodology epoch" – were dominated by a medical/healthcare-based approach, including DRGs, DxCGs, ERGs, etc.

All predictors were with medical terminology and meaning; any additional sources of information mainly stressed the medical information they contained, such as lab data and health risk assessment questionnaires that asked mainly about diseases and drugs. The first HRA questions without direct or indirect medical context were about style of living, including exercise and physical activities, drinking and level of education.

But HRA information is not available in all medical records, and the medical information in them and in lab data is correlated with the diseases and drugs already available in the medical and Rx claims records. There are no new sources of medical information that will add additional predictive power to the currently available medical knowledge. The genomics data is likely the only medical source that has not been explored enough, but it requires significant investment to become accessible – probably within 20 to 30 years. So improvement in healthcare outcome predictive models' accuracy requires new sources of information, sources that are easily available for a huge percentage of the population -- but that are not correlated with already-available medical sources. The choices are:

- socioeconomic data,
- lifestyle data,
- consumer data,
- employment data and
- unstructured data such as text mining and social media analytics.

The advantages compared to other choices are availability, reliability, human experience working with such data and cost. That's why the integration of medical and pharmacy data with socioeconomic data is next logical step of the healthcare outcome predictive modeling industry in its attempt to improve the accuracy of predictions for as many people as possible.

(continued on page 2)

Socioeconomic Data Fuels Deeper Understanding ...continued from page 1

PMN: What are the most common types of socioeconomic data used in healthcare predictive modeling?

OA: We have spent a significant amount of effort testing the validity of data attributes pulled from public records. Key categories that are correlated to health outcomes include level of education, address stability, income, taxes, properties, assets and change in housing costs. Derogatory attributes include arrests, liens, bankruptcies, evictions, accidents, evidence of unpaid financial obligations, high-interest loan applications, relatives' information, business associates' information and information on neighborhood burglaries and other crimes.

It is very important to note that very often the trends in those attributes are more important than the values themselves. For example, a serious reduction in income could be a harbinger of high stress, with all its associated health consequences, and a woman's last name change during the last 24 months could be a sign of a future pregnancy or divorce -- and their resulting stress.

PMN: What are the most common sources of those socioeconomic data?

OA: LexisNexis leverages more than 45 billion records from more than 10,000 different sources, including:

- 17.6 billion consumer records,
- 9.5 billion unique name and address records,
- 4.4 billion property records,
- 46.5 million active US business entities,
- 1.5 billion business contact records,
- 257.6 million unique cell phone numbers,
- 1.5 billion bankruptcy records monitored monthly,
- 4 billion motor vehicle registrations,
- 1 billion vehicle title records and
- 435 million criminal records.

PMN: How customizable to the client is the incorporation of socioeconomic data? Can a company like LexisNexis use just selected portions of its data stores?

OA: That depends on the customer's targets, whether cost, hospitalizations, 30-day readmissions, emergency room visits, medication adherence, motivation or stress. LexisNexis has created the EASE -- for "Enhanced Attribute Selection" -- framework, a methodology for comprehensive evaluation of the predictive power of different attributes with respect to important healthcare outcomes, based on a 24-month medical and pharmacy claim history and more than 400 socioeconomic attributes. There is a lot of useful information for each one of them, and they can be ranked according to their predictive power. The client's target prioritization determines which data subset of attributes will be selected for use. We are under a strict regulatory environment, so the permissible use of this data is very much limited to improving the coordination of proactive care management to improve outcomes and reduce costs.

PMN: What are the most common uses of socioeconomic data in healthcare predictive modeling?

OA: Generally speaking, socioeconomic data is used for predicting health risk in conjunction with -- or in the absence of -- other risk adjustment models, improving the accuracy of existing models, predicting risk for individuals who have no clinical history or overcoming operational challenges posed by bad or missing data, like wrong addresses and phone numbers. There have been many HOPM research studies involving socioeconomic data; they've analyzed the relationship between healthcare targets and potential predictors like employment status, education, car ownership, ethnicity, gender, marital status, insurance and occupational social class. Many studies have investigated the hypothesis that socioeconomic differences in health status change can largely be explained by the higher prevalence of individual health risk behaviors among those of lower socioeconomic position; however, these analyses and predictive models have been done for research purposes. I think now is the time for creating healthcare outcome predictive models that could be applied for population health management using integration of medical and socioeconomic data.

PMN: Please pick a couple use cases and describe how the addition of socioeconomic data enhances the value of existing clinical and claims data.

OA: I'll discuss two recently created healthcare outcome predictive models, the LexisNexis Socioeconomic Risk Score to predict future total cost and the LexisNexis Stress Index to predict/assess future stress severity.

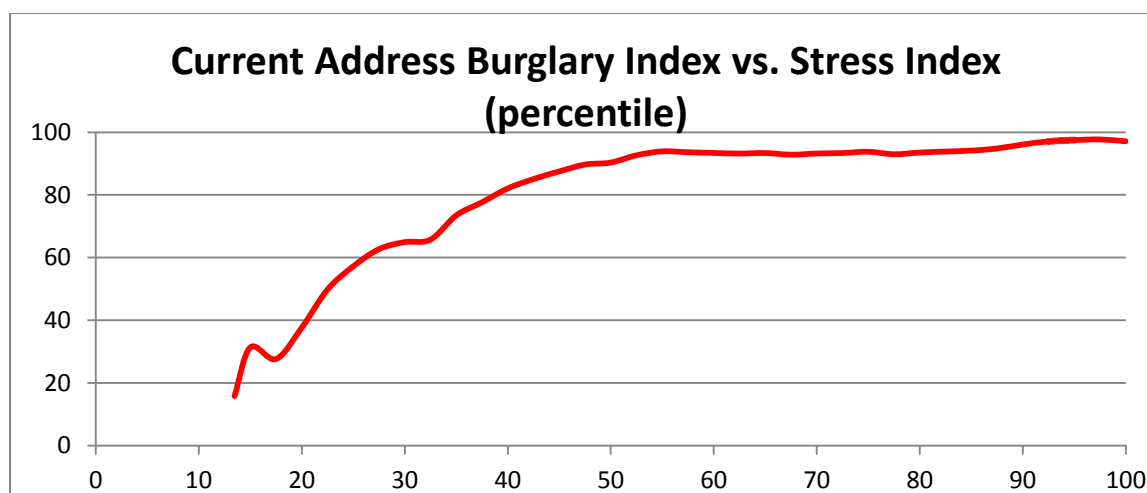
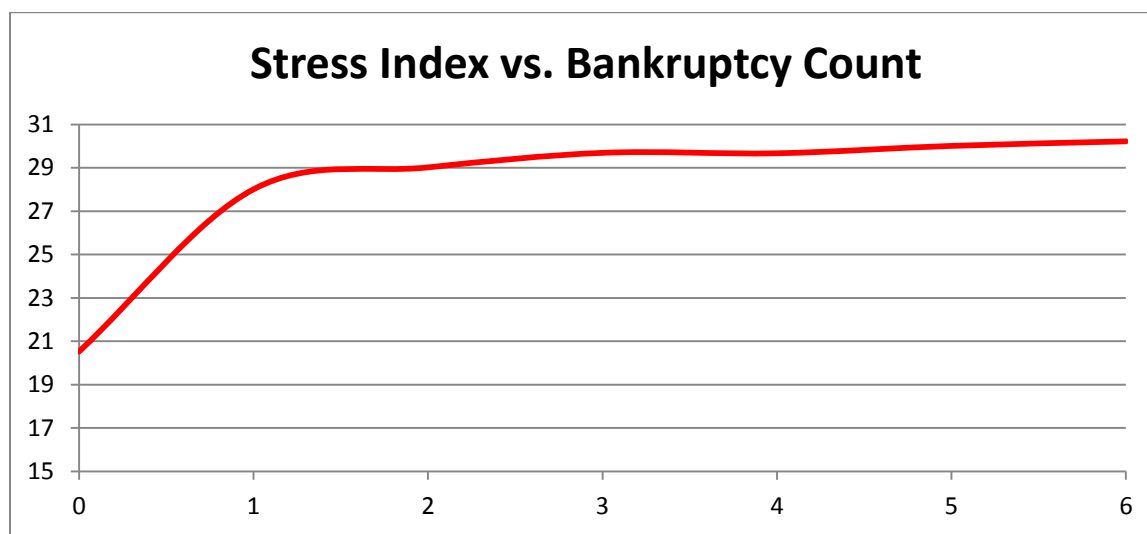
Case 1: Prediction of Future Total Cost. Consider this attribute: Address Recent Economic Trajectory Index, which indicates the property type and value trajectory for an input address. It shows that people who downsize their houses have higher healthcare costs, less education, more applications for high-interest loans and lower income than people who move up. There is no way that trend will be captured by information available in medical and pharmacy claims.

(continued on page 3)

Socioeconomic Data Fuels Deeper Understanding ...continued from page 2

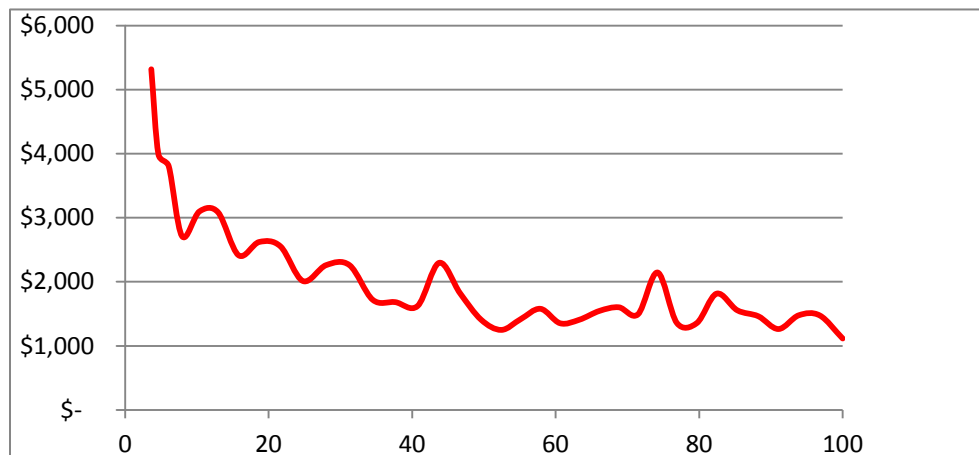
Index	Count	Age	PMPM	Level of Education	Application for high-interest loans	Income
Moved downscale	119,545	49.1	\$480	0.19	0.55	\$37,473
Moved down	465,376	49.6	\$472	0.25	0.24	\$48,243
Stayed	357,048	49.6	\$455	0.33	0.15	\$61,590
Moved up	115,758	51.0	\$449	0.37	0.06	\$87,377
Moved upscale	16,028	52.1	\$459	0.34	0.06	\$104,537
Stayed upscale	8,164	51.8	\$456	0.38	0.05	\$95,637

Case 2: Prediction of Future Stress Severity. More bankruptcies means more stress! More crime in the neighborhood certainly means more stress! And income increases are associated with less stress!



(continued on page 4)

Socioeconomic Data Fuels Deeper Understanding ...continued from page 3



PMN: *How much of a game changer is the addition of socioeconomic data in healthcare predictive analytics?*

OA: The greatest advantage of socioeconomic data is its availability, reliability, human experience working with it and cost. Low correlation with already-available medical data is both an advantage and a disadvantage; however, socioeconomic data reveals some behavioral patterns that give us the opportunity to identify pockets of high health risk “hidden” in medical data. Socioeconomic data could be of great help for new health plan enrollees because for them the only available information is age and gender. Behavioral patterns cannot be identified by an age/gender model, which usually underestimates their health consequences. But those patterns are the “hidden” strength of the socioeconomic data-based predictions.

PMN: *Are there as-yet-untapped sources of socioeconomic data? What’s the holy grail of socioeconomic data sources?*

OA: There probably are other promising untapped sources of socioeconomic data, like social media analytics and text mining. However, I think the industry should remain focused on behavioral pattern definitions tailored to health problems (timing of these patterns is also very important) and interpersonal relations and connections, because the stress of kids, siblings and their parents sometimes could be highly correlated.

PMN: *What’s on the horizon as far as new ways to use socioeconomic data in healthcare predictive modeling? Are there uses that haven’t been perfected yet? Are there uses that need new sources of data to be realized?*

OA: Socioeconomic data should be involved in the prediction of the following healthcare outcomes: stress, motivation, medication adherence, readmissions and disease complications. All are extremely important in the population health management arena, and socioeconomic data could help improve the accuracy of some healthcare outcomes, such as myocardial infarction and intervention plan success.

Contact Asparouhov via Lizzy Feliciano at 407-922-2734 or at leonides.feliciano@lexisnexis.com.