Symptom trends, population-level patient clustering, and adverse event prediction in synthetic electronic health records

Seth Rhoades
Fulgens Consulting, LLC

Key technical elements in this work:

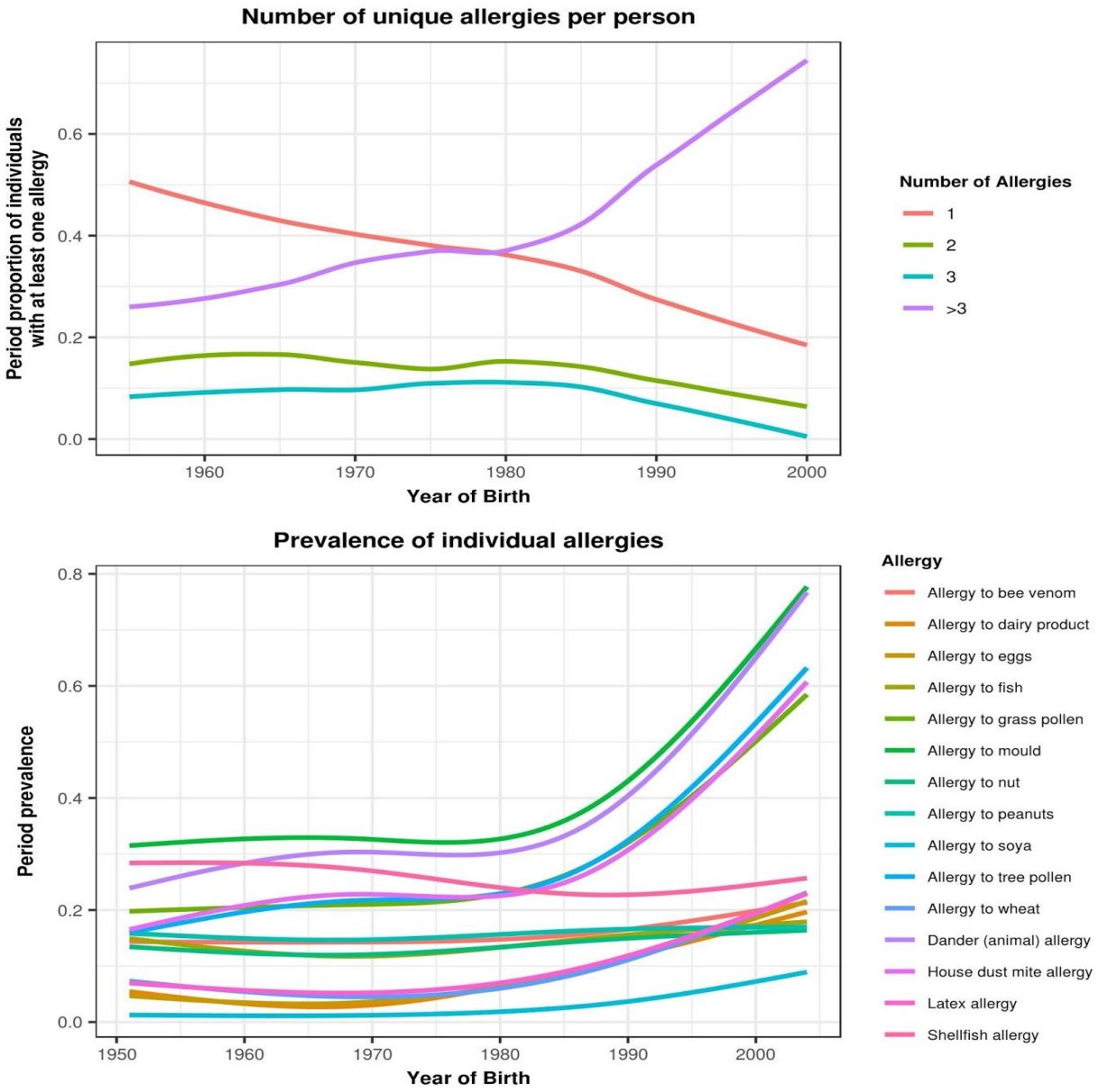
- *Non-linear modeling of temporal health trends*
- *Dimensionality reduction in quantitative population-level data*
- *Identification of key factors which cluster subsets of patients*
- *Event prediction through neural network-driven sequence classification*

** *Code for this analysis is available at https://github.com/fulgensconsulting/DemoProjects/tree/master/synthea* **

Population health is an important area of study for guiding optimal health policies and assessing risks in particularly vulnerable segments of the population. To this end, this report explores the utility of clinical information on large patient populations through synthetic electronic health records derived from the Synthea™ Patient Generator (Walonoski et al., JAMIA, 2017). While the data used in this report is artificial, these analyses can be applied to real-world data to augment our understanding of disease patterns and make meaningful predictions of adverse health events.

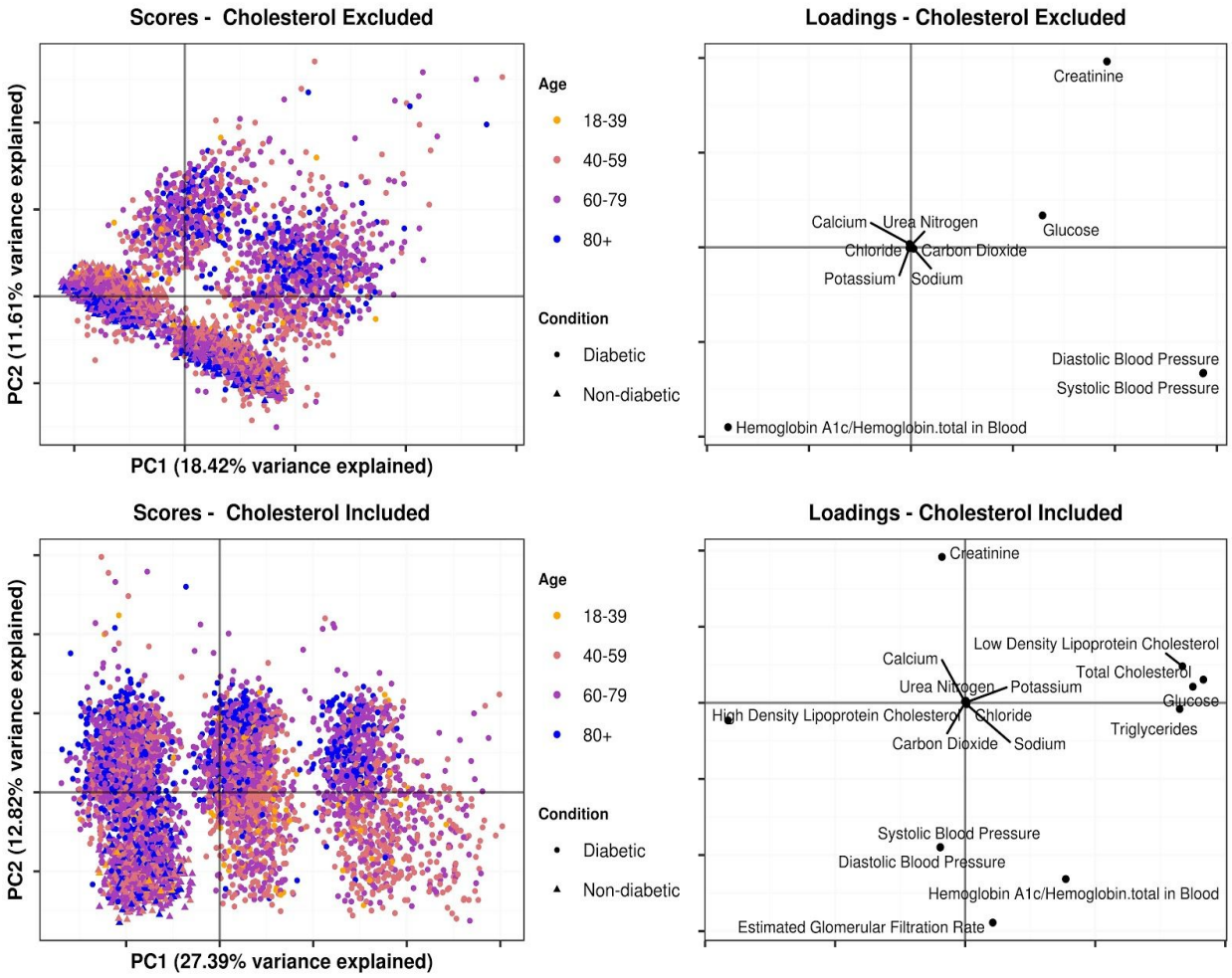
For instance, through the exploration of temporal patterns across clinical histories, this dataset reveals not only an increase of particular allergies in children and adolescents, but a sharp rise in individuals who possess multiple allergies. Subsequently, these findings may highlight the need for a more comprehensive allergy assessment at a young age.

Additionally, objective quantitative measurements can reveal key factors which differentiate subsets of individuals. Principal components analysis (PCA) on high-dimensional data reveals the significance of age, blood pressure, creatinine, and hemoglobin A1c levels to broadly cluster this large population into four groups. While measures of cholesterol and kidney function are not as abundant as blood pressure, adding in these factors to the PCA analysis reveals further clustering within this population. Not surprisingly, these metabolic markers strongly associate with a history of diabetes in these patient's medical records. Thus, objective quantitative measurements provide significant value in determining the likelihood of a patient harboring a given condition even if an individual's clinical history is uncertain.



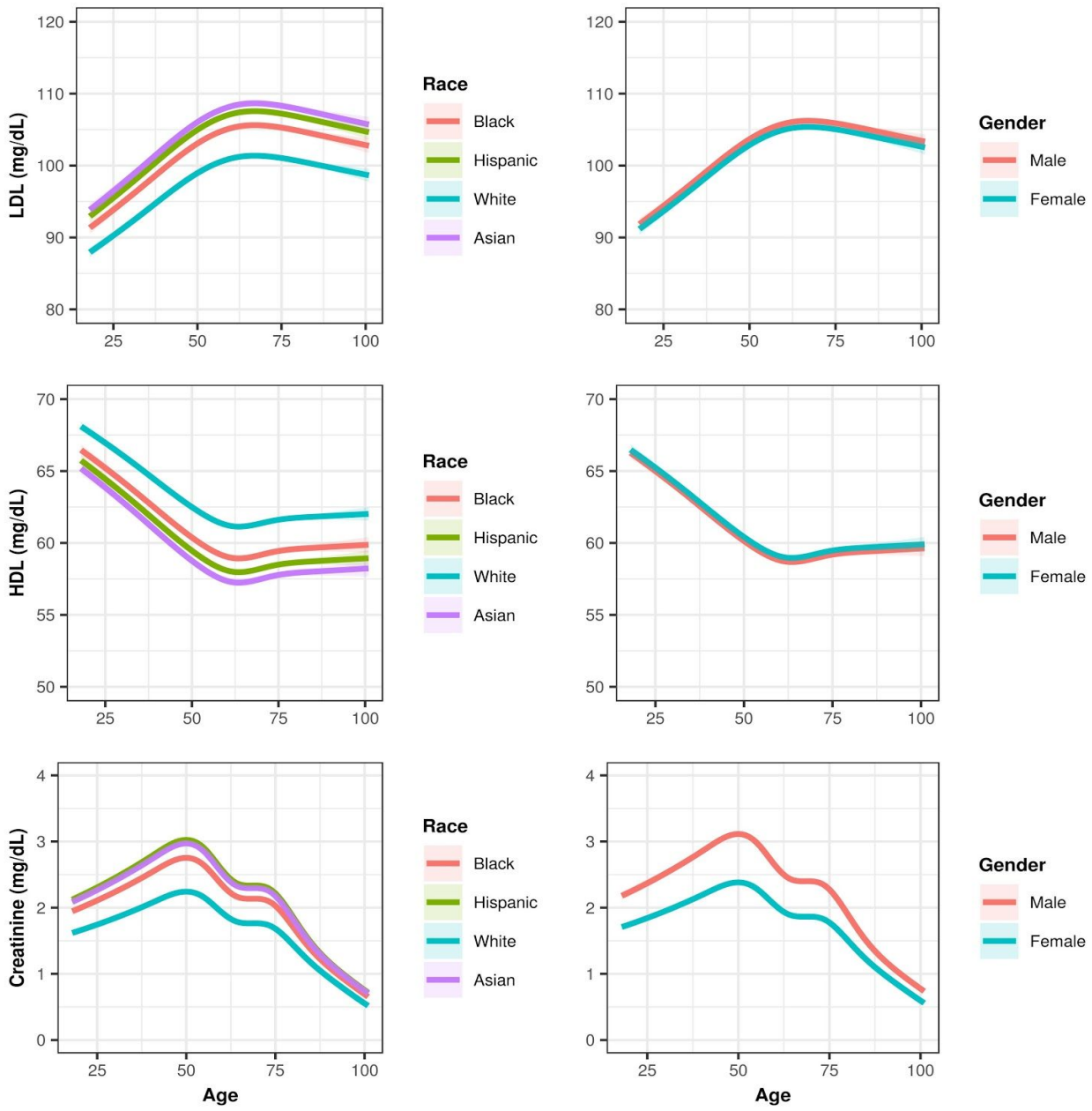
Analysis of allergy period prevalence reveals a rise in individuals born after 1980 who possess more than three allergies (top), driven largely by a subset of particular allergies (bottom).

PCA analysis can generate insights which warrant further investigation. For example, important variables derived from this analysis can be assessed for variation across racial backgrounds and gender. Low-density lipoprotein cholesterol levels are higher in non-white races, which may indicate increased risk of metabolic diseases. Creatinine is markedly different by race and sex, and partly explains the broad clustering by age in the PCA analysis. These additional analyses demonstrate the utility of continual monitoring of key quantitative measurements to provide insight into medical risk across demographic backgrounds.



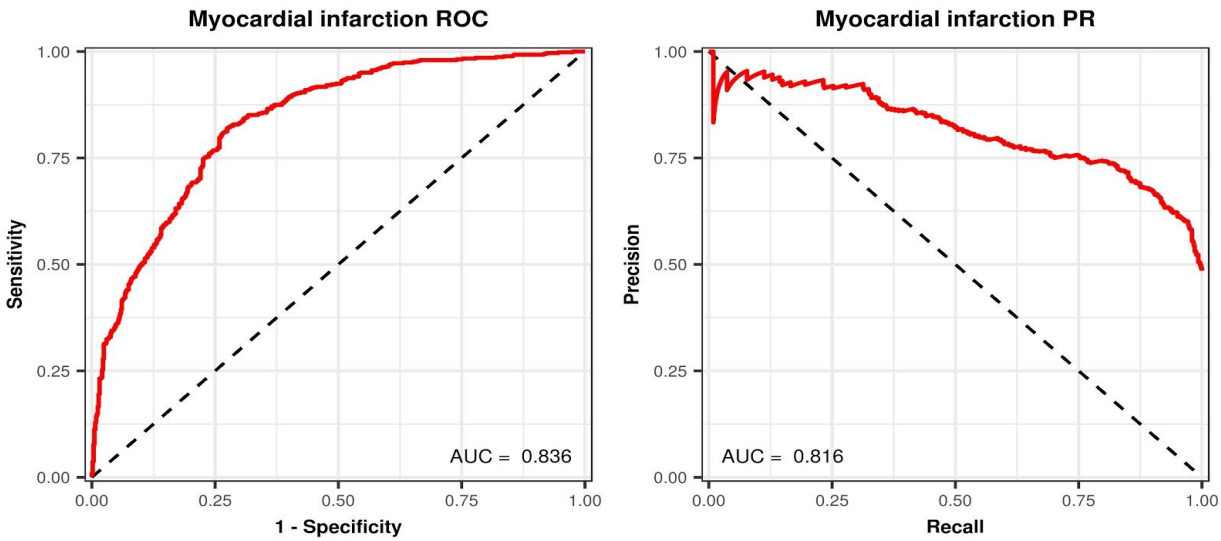
Principal components analysis on broadly accessible quantitative measurements reveals the influence of blood pressure, creatinine, and hemoglobin A1c in separating subsets of patients (top). Other measures, such as cholesterol, are not as abundant in this dataset, but act as additionally important variables in patient clustering (bottom).

Beyond purely observational analysis, predictive modeling on clinical histories can assess risks of future adverse events. Given five-year sequences of multi-modal data types, including demographic variables, geographic location, blood pressure, body mass index, smoking status, procedures, and diagnosis histories, sequence classification can be performed to predict future myocardial infarctions. A deep recurrent neural network model demonstrated 73.4% accuracy on a validation dataset, with receiver operating characteristic and precision-recall area under the curves of 0.836 and 0.816 respectively. This predictive analysis clearly demonstrates the power of advanced statistical modeling on growing longitudinal health information in large populations. Importantly, these models can provide actionable guidance to healthcare providers to both save lives and save costs through measures such as preventative care.



After principal components analysis revealed the influence of cholesterol and creatinine on population-level patient stratification, these variables were modeled to gauge possible risk profiles of metabolic conditions across race, gender, and age.

Collectively, these exploratory analyses demonstrate the utility of large health records for understanding population health, and the power of advanced predictive statistical modeling to predict future events. These approaches highlight some of the many ways in which Fulgens can illuminate actionable insights from complex real-world data.



Receiver operating characteristic (ROC) and precision-recall (PR) curves on myocardial infarction predictions from a deep recurrent neural network sequence classifier on patient histories.

References

Walonoski J, Kramer M, Nichols J, et al. Synthea: An approach, method, and software mechanism for generating synthetic patients and the synthetic electronic health care record. J Am Med Inform Assoc. 2018; 3:230-38.