Specialty Healthcare Benefits Council
Defining Specialty Health

Predictive Analytics: A Case Study in Machine-Learning and Claims Databases

David A. Kvancz, MS, RPh, FASHP; Marcus N. Sredzinski, PharmD; and Celynda G. Tadlock, PharmD, MBA

ABSTRACT

The healthcare industry generates massive collections of data—big data—with the potential to reveal insights into optimizing costs and outcomes if analyzed with the proper tools. Machine learning and predictive algorithms are already in common use in other fields. In healthcare, the largest datasets with the broadest relevance to the US population may reside in claims databases. Analyzing such databases with the latest tools may find rare or hard-to-diagnose diseases that needlessly consume healthcare dollars before proper diagnoses are made.

Objective: This study focused the power of modern analytics on hereditary angioedema (HAE), a single rare disease, because it exhibits features of diseases associated with high costs: rare, hard to diagnose, progressive, and takes a long time from diagnosis to appropriate treatment. Despite the availability of effective therapies, misdiagnoses and underdiagnosis of HAE result in significant burden to the healthcare system.

Methods: A 3-stage process was applied to a claims database to: a) define the characteristics (diagnoses, procedures, therapies, and providers) of patients in the database already being treated for HAE, b) use those characteristics to create a model of patients with HAE, and c) use the model to identify patients with HAE in the database who were not yet diagnosed.

Conclusions: This study successfully demonstrated the ability of this state-of-the-art predictive analysis to find rare-disease patients in a large and complex database. This information could be valuable to claims managers and employers who may realize savings by helping physicians bring these patients to appropriate treatment sooner.

Am J Pharm Benefits. 2016;8(6):214-219

Big data—large pools of diverse data that are collected, stored, and analyzed to reveal unexpected patterns and relationships—has rapidly evolved to shape and inform nearly all sectors of the global economy.[1] Ultimately, big data seeks to play a useful economic role by revealing the potential value hidden in this information. The development of tools capable of extracting value from these massive collections of information has made big data relevant to all sectors of the market. Consumers and providers of products and services, as well as governments and regulators, all stand to benefit from the insights emerging from the new science of big data.[1-3]

Healthcare stands out as a sector with a great deal to gain from the potential of big data. A 2011 McKinsey Global Institute report estimated that if the US healthcare system could successfully apply big data to drive efficiency and quality, the annual potential realized value could be more than $300 billion, two-thirds of which would arise from an 8% reduction in expenditures.[1]

Experts have recommended that healthcare adopt big-data approaches, and such US organizations as the Department of Veterans Affairs and Kaiser Permanente, the integrated managed-care consortium, have implemented innovative pilot programs, many of which utilize clinical data from electronic health records (EHRs) to identify cost-savings opportunities from clinical practice patterns.[1,4,5] Broader adoption of big-data analytics will continue to grow as healthcare organizations strive to comply with the accountability requirements of the Patient Protection and Affordable Care Act.[4]

The lagging rate of adoption of EHRs in the United States has been a barrier to fully leveraging the potential of this data. However, even as EHR implementation has progressed, access to this information continues to stand in the way of high-quality research regarding treatment effectiveness and cost efficiency.[6-8] Data-sharing and transparency guidelines have only recently been put forward, and policies regarding ethics and regulation still need to be defined by participating institutions.[6] One possible way to circumvent the issues raised by using and sharing EHRs is

to use insurance claims data composed of de-identified diagnosis-related details and payment information. The size and national scope of a claims database are also advantages over the typically local or regional nature of EHRs when trying to identify opportunities within the data. One such claims database aggregated data for more than 170 million US patients from 2006 to 2014.[9] Claims datasets allow for in-depth assessment of health and quality outcomes when analyzed with tools capable of handling datasets of this size.

Relevant predictive algorithms and machine-learning techniques designed to handle massive datasets have been available for years, but their applicability to healthcare has not been recognized until relatively recently.[3] For example, predictive analytics designed to assess risks and to model likely outcomes from disparate data types (geospatial, text reports, equipment inventories, etc) are used by the military to enhance operational efficiencies and have applicability to many other fields, such as criminal investigation, business, and healthcare.[10-13] Predictive systems, driven by machine-learning techniques that evolve based on empirical data, are ideal tools for recognizing the patterns obscured by the volume of insurance claims data.[1,4]

Identification of patients with rare diseases within the claims database may offer an opportunity to uncover significant value. Rare diseases, any one of which affects fewer than 200,000 persons, currently affect about 10% of the US population, or more than 32 million individuals.[14] For example, hereditary angioedema (HAE) is a potentially fatal rare disease with a prevalence of 1 in 10,000 to 50,000 in the United States.[15,16] Individuals with the disease may be misdiagnosed for as long as 8 years.[15,16] This genetic, autosomal-dominant disease causes recurrent, painful attacks of subcutaneous and submucosal swelling of the skin, gastrointestinal tract, and larynx.[17] Although not associated with hives, the skin swelling of HAE often leads to misdiagnosis as allergic reaction.[18] Swelling of the gastrointestinal tract produces pain, distension, nausea, vomiting, and diarrhea.[15] Because patients may experience abdominal symptoms for many years before manifesting the characteristic subcutaneous swelling of HAE, patients often undergo inappropriate surgical and medical treatment for any of a wide range of mistaken diagnoses, including acute abdomen, biliary colic, hepatitis, regional enteritis, pancreatitis, cholecystitis, nephrolithiasis, pyelonephritis, ruptured ovarian cyst, intestinal obstruction, duodenal ulcer, and ulcerative colitis.[15] Edema of the larynx may lead to suffocation and death.[18]

Despite the availability of effective therapeutic approaches that address acute treatment, short-term prophylaxis, and maintenance therapy, misdiagnoses and underdiagnosis of HAE result in significant burden to the healthcare system.[15,19] Between 2006 and 2007, US patients with HAE who were misdiagnosed accounted for 5040 emergency department (ED) visits, 41% of which resulted in hospitalization.[15] At an average cost of $1479 per ED visit and an average $22,728 for a 5-day hospitalization, significant cost savings might be realized through more prompt and effective diagnosis and management of these patients.[15]

The study presented here was designed to demonstrate the ability of a modern, automated machine-learning system to discover undiagnosed rare-disease patients in a claims database. The main contributions of this research are to show how the adaptation of state-of-the-art technologies in big-data analytics, information theory, and machine learning can create a seamlessly integrated framework for database analysis and to demonstrate how this technology could be put to practical use using insurance claims, rather than EHR data, to find undiagnosed rare-disease patients. The case presented here focused on HAE.[16]

PRACTICAL IMPLICATIONS

An analytical tool with the flexibility to be applied to a variety of data sources and specifically identify very small patient subpopulations has the potential to be a powerful force in the evolving healthcare landscape.

- Using such a tool, payers may realize cost savings from identifying patients with costly conditions as early as possible and take steps to ensure that their care is managed appropriately.
- Physicians would be better informed about how to manage these patients, and patients would receive the care best suited for their needs.
- These analytical techniques could apply to open and closed healthcare systems, both large and small.

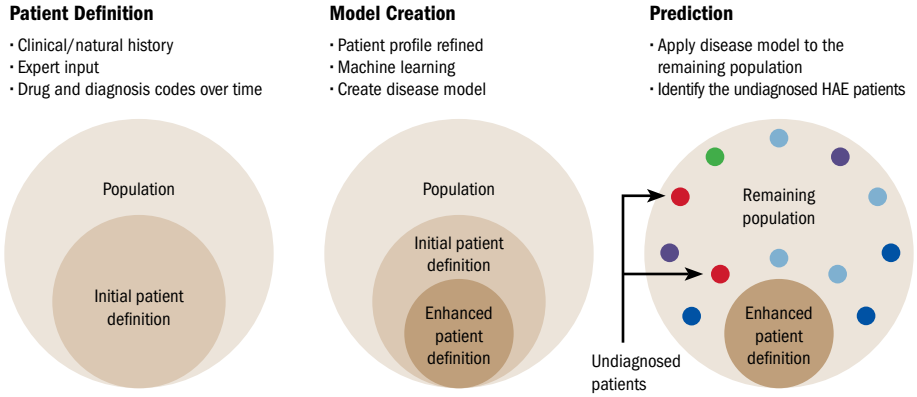
METHODS

The population for this analysis was extracted from a database of de-identified patient claims data acquired from Truven Health Analytics (MarketScan claims data). This claims database contains health insurance claims data for more than 170 million unique lives covering 2006 through 2014. A 3-stage process was employed to discover patients with HAE within this database who had not been diagnosed (Figure 1).

Stage 1: Patient Definition

To study the characteristics of HAE sufferers and to identify their statistical "signature," the first step was to find a group of patients in the database who definitely had HAE. Diagnosis codes from *International Classification of Diseases, Ninth Revision, Clinical Modification* (*ICD-9-CM*) alone may not be fully reliable: a code may be used

Figure 1. **Three-stage Discovery Process for Identifying Undiagnosed Hereditary Angioedema Patients in an Insurance Claims Database**



for billing purposes without official diagnosis; old codes may be used even after new, more specific codes become available; an *ICD-9-CM* code sometimes represents a group of diseases; and data entry errors could occur. Consulting physicians and pharmacy experts for this study agreed that patients prescribed 1 of the 4 HAE-specific drugs available in the United States were, without doubt, patients with HAE. The 4 drugs were Cinryze (C1 esterase inhibitor [human]; Shire), Firazyr (icatibant; Shire), Berinert (C1 esterase inhibitor [human]; CSL Behring), and Kalbitor (ecallantide; Shire). Thus, patients identified in the database as being prescribed 1 or more of these 4 drugs formed the population of "index HAE patients."

Stage 2: Model Creation

To identify which features or combination of features are most statistically relevant for differentiating HAE from non-HAE patients, an information-theoretic concept of mutual information (MI) was utilized to determine the differentiating features. MI is a measure of how much information about one set of data can be determined from another set of data.[20] In this analysis, the features with higher MI values were likely to be more informative for discriminating HAE from non-HAE patients. After the MI of individual features or their combinations was computed, the process of feature selection began. The goal of feature selection was to define the smallest subset of features that collectively contain most of the mutually shared information and thus most clearly define the characteristics of the patient with HAE. Machine-learning algorithms drove the analysis of feature selection that created a model of HAE. Thus, the model consisted of the fewest possible and simultaneously most differentiating characteristics of patients with HAE, resulting in an enhanced patient definition.

Stage 3: Prediction

Once a model of the characteristics of the patient with HAE was determined from the index patients with HAE, the remaining population of patients in the data set was scored by the model to find undiagnosed patients. For every remaining patient in the data set, the first step in scoring was to compute the features that did not appear in the set of index patients with HAE. Each patient's features were input to the HAE model, which produced a numerical score. This score represented the likelihood that the patient had undiagnosed HAE, and patients were ranked from most likely to least likely to have the condition.

RESULTS

Stage 1: Patient Definition

Searching the 2006-2014 MarketScan database for all patients prescribed two C1 inhibitors [human] (Cinryze, Berinert), incatibant (Firazyr), and ecallantide (Kalbitor) revealed 1002 index patients with HAE.

Stage 2: Model Creation

The histories of the patients identified in Stage 1 were analyzed to determine the diagnostic, procedural, therapeutic, and healthcare provider characteristics prior to receiving definitive treatment for HAE. By comparing the characteristics of patients with HAE with those of demographically matched non-HAE patients, machine-learning algorithms selected the characteristics that were most descriptive and predictive of eventual HAE diagnosis by a physician. These characteristics were identified and refined, forming the enhanced patient-definition characteristics listed in the Table encompassing diagnosis, procedures, therapies, and providers. Note that some nonadjacent lines of descriptive text appear identical within the table. These are associated with different *ICD-9-CM* codes depending on the provider's level of involvement

with the patient; thus, they appear more than once in the table with different ranking.

Stage 3: Prediction

A model of the HAE patient's history and profile, based on the enhanced patient definition determined in Stage 2, was applied to the remaining population of patients in the database. With the prediction classifier set to a detection probability ≥0.8 in Stage 3 of this analysis, applying the model to the remaining population indicated 5511 potentially undiagnosed patients with HAE.

Although the data in the database is de-identified, the patient information in the database is linked to metropolitan statistical areas (MSAs) to understand the geographic distribution of the information. The Office of Management and Budget defines MSAs for use by federal statistical agencies.[21] The distribution of the predicted HAE patients across the United States is depicted in the map in Figure 2.

DISCUSSION

The major contributions of this study relate to the identification of patients with a rare disease, the analytic method used, and the source of data used. Together, these elements define a new environment in which payers, physicians, and patients may benefit from the value still locked away in big healthcare data.

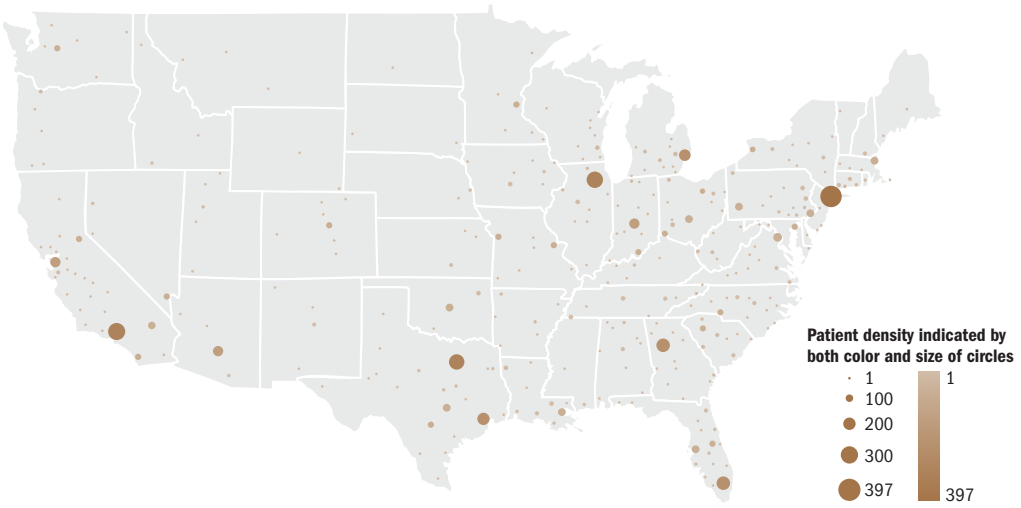
As healthcare gradually embraces the value of data analytics, it still struggles with overcoming access and transparency issues with regard to using patient records.[6-8] As a result, many of the efforts to apply predictive analytics are directed at the EHRs of a single institution or network. This produces results that may have limited relevance to other health provider systems, are based on a limited population size, and often focus on providing rapid feedback to alert the healthcare provider of potential care issues. The frequency and ubiquity of these alerts, often with limited practical value, has been known to produce "alert fatigue," which results in some healthcare providers ignoring these warnings, thus further diminishing the usefulness of these predictive analytics. As a result, the goal of these analytics to provide decision-making mechanisms that maximize the value of medical care are not fully realized. By utilizing a de-identified claims database compliant with the Health Insurance Portability and Accountability Act of 1996, barriers to transparency and sharing are overcome. The size of the database—more than 170 million patients in this study—ensures confidence in the relevance of the outcomes to the US population and in the significance of the results. Analyzing a claims database rather than EHRs may allow some payers to more effectively focus an analysis on

Table. The Top 10 Diagnostic, Procedural, Therapeutic, and Healthcare Provider Characteristics Most Predictive of Heredity Angioedema Diagnosis[a]

Diagnosis	
1	Allergic reactions
2	Swelling, mass, or lump in head and neck
3	Routine general medical examination at a healthcare facility
4	Immunizations and screening for infectious disease
5	Other screening for suspected conditions (not mental disorders or infectious disease)
6	Edema
7	Abdominal pain, unspecified site
8	Other upper respiratory disease
9	Unspecified symptom associated with female genital organs
10	Chronic vascular insufficiency of the intestine
Procedures	
1	Office or other outpatient visit for the evaluation and management of an established patient
2	Other laboratory
3	Office or other outpatient visit for the evaluation and management of an established patient
4	Laboratory: chemistry and hematology
5	Other therapeutic procedures
6	Pathology
7	Other diagnostic radiology and related techniques
8	Microscopic examination (bacterial smear, culture, toxicology)
9	Office or other outpatient visit for the evaluation and management of an established patient
10	Nonoperative urinary system measurements
Therapy	
1	Androgens and combinations
2	Blood derivatives
3	Androgens and combinations
4	Unspecified agents
5	Sympathomimetic agents
6	Adrenals and combinations
7	Analgesics/antipyretics; opiate agonists
8	Antibiotics: penicillins
9	Antibiotics: erythromycin and macrolide
10	Analgesics/antipyretics; nonsteroidal anti-inflammatory drugs
Provider	
1	Outpatient hospital
2	Office
3	Independent laboratory
4	Emergency department (hospital)
5	Inpatient hospital
6	Independent clinic
7	Patient home
8	Outpatient (not elsewhere classified)
9	Ambulatory surgical center
10	Ambulance (land)

[a]Text descriptions for any characteristic that appear identical are actually associated with different *ICD-9-CM* codes depending on the provider's level of involvement with the patient.

Figure 2. **Distribution of Potential Heredity Angioedema Patients Within Metropolitan Statistical Areas, 2006-2014 (N = 5511)**



uncommon disease states to answer questions about the best ways to contain costs while maximizing care options for the population they serve.

An opportunity arose to affirm the computed prediction of undiagnosed patients. An update to the MarketScan database covering the 11 months from January 2015 through November 2015 was scanned for further information about patients with potentially undiagnosed HAE. A total of 888 of the predicted 5511 undiagnosed patients were found to have new healthcare information during those 11 months. Of those 888 potential HAE patients, 14 had new claims data codes for HAE, thus affirming the relevance of the computed predictive model. Although this does not constitute statistically rigorous validation of the model, confirmation of diagnosis in these 14 patients predicted to have HAE suggests the potential power of this model to have an impact on cost and outcome management for rare and hard-to-diagnose diseases.

This study focused the power of state-of-the-art analytics on a single rare disease, HAE, because HAE shares many disease features associated with high costs: rare, hard to diagnose, progressive, and takes a long time from diagnosis to appropriate treatment. During the 8 years it may take to diagnose a patient with HAE, patients frequently visit EDs, are admitted for hospital stays, and often receive inappropriate ans expensive procedures.[15] Earlier diagnosis and treatment would remove patients from the cycle of high-cost, ineffective treatment that drives them back for more of the same, thus reducing waste and improving patient outcomes. With 3.2 million potential patients with rare diseases in the United States, predictive analytics applied to claims databases to identify them could open the door for payers to help physicians maximize outcomes and value in the care of these patients.[14]

Machine-learning algorithms used in this study have crossed over from other disciplines, such as defense and business, that are already demonstrating the flexibility and adaptability inherent in their design.[10-13] This study successfully demonstrated the ability of this state-of-the-art predictive analysis to find rare-disease patients in a large and complex insurance database. An analytical tool with the flexibility to be applied to a variety of data sources and to specifically identify patient subpopulations of interest to payers or healthcare institutions, has the potential to be a powerful force in the evolving healthcare landscape. Using such a tool, payers may realize cost savings from identifying patients with costly conditions and from taking steps to ensure that their care is managed appropriately. Physicians may be better informed about how to manage these patients, and patients would receive the care best suited for their needs.

The techniques used in this analysis could apply to open and closed healthcare systems, both large and small. Large healthcare systems that invest in state-of-the-art predictive analytics would have tools at the ready to answer critical questions about how their patient-population needs are being met and how their costs are allocated. More importantly, such tools could provide insight into what might be done to meet patients' changing needs and respond efficiently to the demands of the evolving managed care system. There is potential for smaller regional systems and individual health plans or employers to apply lessons learned from published analyses of larger systems to guide examination of their own data. Applying the techniques used here to other diseases that are rare, hard to diagnose, progressive, and take a long time from diagnosis to appropriate treatment has the potential to benefit payers, physicians, and patients in the accountable care environment.

Limitations

Certain limitations to this analysis should be considered. Despite being generally representative of the US population, the MarketScan database is composed of data from a subset of the US population and thus is not a random sample.[9,22] The data come mostly from large employers, so medium- and small-firm data are not represented.[22] Administrative claims data typically contain some coding inaccuracies and missing data, which might result in misclassification or other bias. Additionally, although self-validating cross-checks were incorporated as part of developing the analytical model, real-world validation of the identification of these undiagnosed patients is not complete.

CONCLUSIONS

This analysis successfully demonstrated the ability of this state-of-the-art predictive analysis to find potential rare-disease patients in a large and complex database. Machine-learning techniques applied to a de-identified claims database are clearly capable of identifying these undiagnosed and inappropriately treated patients. This information could be valuable to claims managers and employers who may realize savings by helping physicians bring these patients to appropriate treatment sooner. The potential exists to apply this technique to other diseases that are rare, hard to diagnose, progressive, and may take a long time from diagnosis to appropriate treatment. It is a lesson for managed care organizations of all types that new data analysis and patient differentiation techniques are applicable to the patient populations they manage. It is time for healthcare to join other data-intensive industries in embracing technologies that reveal the value in the largest asset they manage: information.

Author Affiliations: Strategic Client Relationships; Visante, Inc (DAK), St. Paul, MN; Medical Security Card Co (MNS), Tucson, AZ; Clinical, Product, and Customer Experience, Aetna Pharmacy (CGT), Atlanta, GA.

Source of Funding: Funded through an unrestricted educational grant from HVH Patient Precision Analytics (HVH), LLC.

Author Disclosures: The authors report no relationship or financial interest with any entity that would pose a conflict of interest with the subject matter of this article.

Authorship Information: Concept and design (DAK, MNS, CGT); analysis and interpretation of data (DAK, MNS, CGT); drafting of the manuscript (DAK, MNS, CGT); critical revision of the manuscript for important intellectual content (DAK, MNS, CGT); supervision (DAK)

Address correspondence to: David A. Kvancz, MS, RPh, Visante, Inc, 101 East Fifth Street, #2220 St. Paul, MN 55101. E-mail: dkvancz@Visanteinc.com

REFERENCES

1. Manyika J, Chui M, Brown B, et al; McKinsey Global Institute. Big data: the next frontier for innovation, competition, and productivity. McKinsey & Company website. http://www.mckinsey.com/business-functions/business-technology/our-insights/big-data-the-next-frontier-for-innovation. Published May 2011. Accessed June 22, 2016.

2. McAfee A, Brynjolfsson E. Big data: the management revolution. *Harvard Business Review* website. https://hbr.org/2012/10/big-data-the-management-revolution. Published October 2012. Accessed June 22, 2016.

3. Naidus E, Celi LA. Big data in healthcare: are we close to it? *Rev Bras Ter Intensiva*. 2016;28(1):8-10. doi:10.5935/0103-507X.20160008.

4. Bates DW, Saria S, Ohno-Machado L, Shah A, Escobar G. Big data in health care: using analytics to identify and manage high-risk and high-cost patients. *Health Aff (Millwood)*. 2014;33(7):1123-1131. doi:10.1377/hlthaff.2014.0041.

5. Parikh RB, Obermeyer Z, Bates DW. Making predictive analytics a routine part of patient care. *Harvard Business Review* website. https://hbr.org/2016/04/making-predictive-analytics-a-routine-part-of-patient-care. Published April 21, 2016. Accessed June 28, 2016.

6. Amarasingham R, Audet AM, Bates DW, et al. Consensus statement on electronic health predictive analytics: a guiding framework to address challenges. *EGEMS (Wash DC)*. 2016;4(1):1163. doi:10.13063/2327-9214.1163.

7. Using real-world evidence to accelerate safe and effective cures: advancing medical innovation for a healthier America. Bipartisan Policy Center website. Published June 2016. Accessed June 27, 2016.

8. Doshi JA, Hendrick FB, Graff JS, Stuart BC. Data, data everywhere, but access remains a big issue for researchers: a review of access policies for publicly-funded patient-level health care data in the United States. *EGEMS (Wash DC)*. 2016;4(2):1204. doi:10.13063/2327-9214.1204.

9. MarketScan Research Databases. Truven Health website. http://truvenhealth.com/your-healthcare-focus/analytic-research/marketscan-research-databases. Accessed June 28, 2016.

10. Klein A. Police enlist war tech in crime fight. *Washington Post* website. https://www.washingtonpost.com/local/police-enlist-war-tech-in-crime-fight/2013/02/18/0a9e18e2-6bc6-11e2-ada0-5ca5fa7ebe79_story.html. Published February 18, 2013. Accessed June 29, 2016.

11. Ward MJ, Marsolo KA, Froehle CM. Applications of business analytics in healthcare. *Bus Horiz*. 2014;57(5):571-582. doi:10.1016/j.bushor.2014.06.003.

12. Wood C. How does the military use big data? Emergency Management website. http://www.emergencymgmt.com/safety/Military-Use-Big-Data.html. Published January 6, 2014. Accessed June 29, 2016.

13. Custom Strategies/IBM Government Analytics Forum. Putting predictive analytics to work for the army—an executive perspective. Government Executive website. http://www.govexec.com/govexec-sponsored/2015/04/putting-predictive-analytics-work-army-executive-perspective/111406/. Published April 30, 2015. Accessed June 29, 2016.

14. Rare diseases: facts and statistics. Global Genes website. http://globalgenes.org/rare-diseases-facts-statistics/. Published January 1, 2012. Accessed June 2, 2016.

15. Ali MA, Borum ML. Hereditary angioedema: what the gastroenterologist needs to know. *Clin Exp Gastroenterol*. 2014;7:435-445. doi:10.2147/CEG.S50465.

16. Lumry WR, Castaldo AJ, Vernon MK, Blaustein MB, Wilson DA, Horn PT. The humanistic burden of hereditary angioedema: impact on health-related quality of life, productivity, and depression. *Allergy Asthma Proc*. 2010;31(5):407-414. doi:10.2500/aap.2010.31.3394.

17. Riedl M. Recombinant human C1 esterase inhibitor in the management of hereditary angioedema. *Clin Drug Investig*. 2015;35(7):407-417. Review. doi:10.1007/s40261-015-0300-z.

18. Agostoni A, Aygören-Pürsün E, Binkley KE, et al. Hereditary and acquired angioedema: problems and progress: proceedings of the third C1 esterase inhibitor deficiency workshop and beyond. *J Allergy Clin Immunol*. 2004;114(suppl 3):S51-S131. doi:10.1016/j.jaci.2004.06.047.

19. Gómez-Traseira C, Pérez-Fernández E, López-Serrano MC, et al. Clinical pattern and acute and long-term management of hereditary angioedema due to C1-esterase inhibitor deficiency. *J Investig Allergol Clin Immunol*. 2015;25(5):358-364.

20. Ross BC. Mutual information between discrete and continuous data sets. *PloS One*. 2014;9(2):e87357. doi:10.1371/journal.pone.0087357.

21. Metropolitan and micropolitan statistical areas main. US Census Bureau website. http://www.census.gov/population/metro/. Updated July 2015. Accessed September 12, 2016.

22. Hansen LG, Chang S. Health research data for the real world: The MarketScan Databases. Truven Health website. http://truvenhealth.com/portals/0/assets/PH_11238_0612_TEMP_MarketScan_WP_FINAL.pdf. Accessed September 9, 2016. ajpb