Novel Predictive Models for Metabolic Syndrome Risk: A "Big Data" Analytic Approach

Gregory B. Steinberg, MB, BCh; Bruce W. Church, PhD; Carol J. McCall, FSA, MAAA; Adam B. Scott, MBA; and Brian P. Kalis, MBA

ABSTRACT

Objectives

We applied a proprietary "big data" analytic platform—Reverse Engineering and Forward Simulation (REFS)—to dimensions of metabolic syndrome extracted from a large data set compiled from Aetna's databases for 1 large national customer. Our goals were to accurately predict subsequent risk of metabolic syndrome and its various factors on both a population and individual level.

Study Design

The study data set included demographic, medical claim, pharmacy claim, laboratory test, and biometric screening results for 36,944 individuals. The platform reverse-engineered functional models of systems from diverse and large data sources and provided a simulation framework for insight generation.

Methods

The platform interrogated data sets from the results of 2 Comprehensive Metabolic Syndrome Screenings (CMSSs) as well as complete coverage records; complete data from medical claims, pharmacy claims, and lab results for 2010 and 2011; and responses to health risk assessment questions.

Results

The platform predicted subsequent risk of metabolic syndrome, both overall and by risk factor, on population and individual levels, with ROC/AUC varying from 0.80 to 0.88. We demonstrated that improving waist circumference and blood glucose yielded the largest benefits on subsequent risk and medical costs. We also showed that adherence to prescribed medications and, particularly, adherence to routine scheduled outpatient doctor visits, reduced subsequent risk.

Conclusions

The platform generated individualized insights using available heterogeneous data within 3 months. The accuracy and short speed to insight with this type of analytic platform allowed Aetna to develop targeted cost-effective care management programs for individuals with or at risk for metabolic syndrome.

Am J Manag Care. 2014;20(6):e221-e228

The growing prevalence of metabolic syndrome in the United States, and globally, is alarming. Metabolic syndrome is generally defined as having three or more of five common biological abnormalities out of range: waist circumference, blood pressure, elevated triglycerides, low high density lipoproteins (HDL),and increased insulin resistance. Analysis[1] suggests that almost one-third of US adults, or approximately 80 million people, meet the Adult Treatment Panel III criteria for metabolic syndrome, with prevalence increasing significantly with age and body weight.[2] An additional 45%, or approximately 104 million people, have 1 or 2 risk factors for developing metabolic syndrome.

These trends have profound clinical and financial implications. Individuals with metabolic syndrome are twice as likely to develop cardiovascular disease and 5 times more likely to develop diabetes mellitus, both of which mean higher than average annual healthcare costs. Workplace participation and productivity of individuals with metabolic syndrome are also negatively impacted.[3]

Health insurance companies have large quantities of data relevant to metabolic syndrome, including demographic data, diagnosis and procedure claim data, lab results, prescription data, and care management program data. Using "big data analytics" to interrogate large, complex data sets can generate meaningful insights about individuals with or at risk of developing metabolic syndrome.

We applied a proprietary "big data" analytic platform—Reverse Engineering and Forward Simulation (REFS)—to the data set of 1 of Aetna's larger nationwide retail customers and calculated:

- The subsequent risk of metabolic syndrome, both overall and by metabolic syndrome risk factor, at both a population and individual level
- The impact of incremental changes in risk factors on the overall subsequent risk of metabolic syndrome and on costs

Take-Away Points

- Health insurance companies have large quantities of data relevant to predicting onset of conditions such as metabolic syndrome, including demographic, diagnosis and procedure claim data, lab results, and prescription and care management program data.
- The platform allows users to interrogate such large, complex data sets and generate meaningful insights within months about individuals and populations at risk, and for a fraction of the cost of clinical trials and traditional analysis.
- The speed-to-insight possible with this new approach allowed Aetna to design and launch customized interventions to improve health outcomes of the affected population and start quantifying returns on its program investment.

- The impact of adherence to medications and to routine, scheduled outpatient doctor visits on the subsequent risk of metabolic syndrome.

Big data analytic techniques of this type rapidly yiled insights that support data-driven targeted interventions for people with or at risk of developing metabolic syndrome. Aetna is currently piloting an intervention program based upon the results.

METHODS

The REFS platform is best used to analyze and simulate large, dynamic, multisource data sets. The platform learns by reverse engineering ensembles of models that represent the diversity of processes consistent with the data and then simulating nonparametric knowledge representations to generate accurate, granular group and individual predictions that are both actionable and generalizable. Accurate insights from available data can be generated within a few months, and new data easily integrated. The speed-to-insight allows care providers to develop effective therapeutic programs and interventions quickly and cost-effectively, ultimately lowering the cost to serve the affected populations.

Data Sources

Data for this study were gathered from:

- Insurance eligibility records
- Medical claims records
- Pharmacy claims records
- Comprehensive Metabolic Syndrome Screening (CMSS) results
- Laboratory test results
- Health risk assessment (HRA) responses

Study Population

The CMSS results provided the core outcome variables for the study, and measured each of the 5 metabolic syndrome factors (including systolic and diastolic blood pressure). Screenings were conducted twice: once at the beginning of 2011 and again in early 2012, for an initial cohort of 59,605 people. We then restricted the study to participants for whom we had: complete coverage records from January 1, 2010, through December 31, 2011; complete data from medical claims, pharmacy claims, or test lab results for 2010 and 2011; and valid responses to a small set of HRA questions. This resulted in a study population of 36,944, which was then randomly assigned to either an 80% training set (N = 29,527) or a 20% test set (N = 7417). The study population metabolic syndrome risk and medical cost profile is found in Figure 1. Additional demographic detail is found in eAppendix Figure 1.

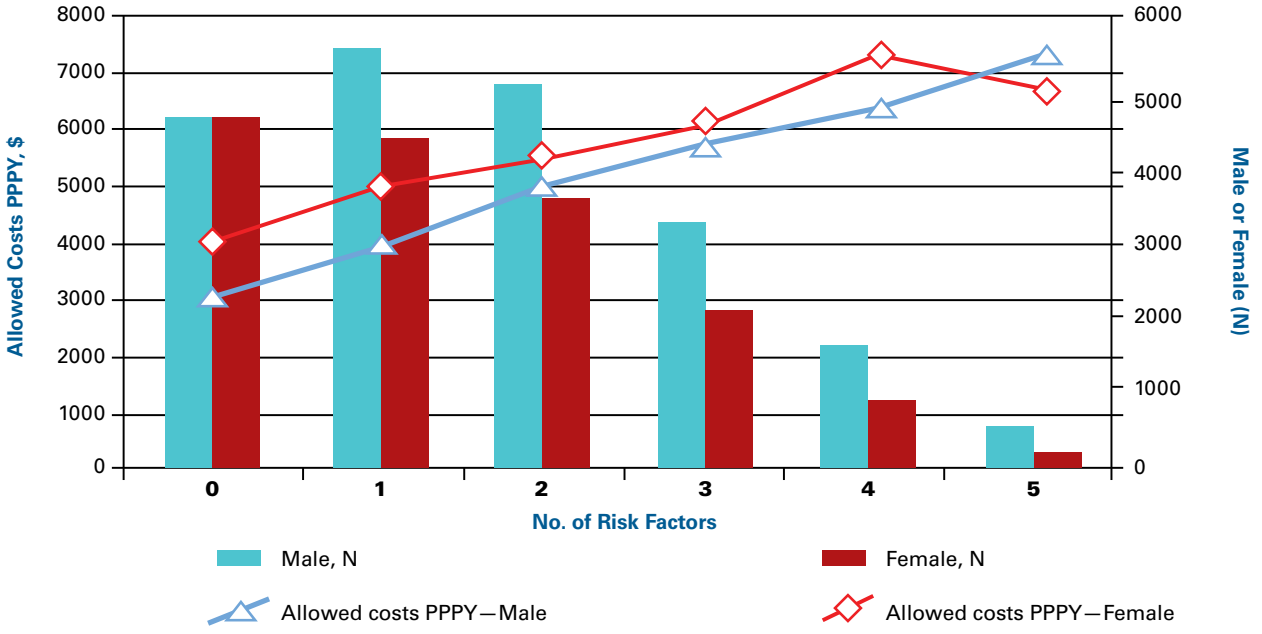
Variable Creation and Definitions

The 4291 variables in the analysis spanned 6 different data categories. The specific breakdown of data categories is found in eAppendix Table 1. Continuous variables were discretized into ranges in preparation for modeling with multivariate categorical models. The ranges of the CMSS factors were constructed from metabolic syndrome out-of-range boundaries and other clinically relevant boundaries.

Demographics captured 5 dimensions in addition to gender: age, body mass index (BMI), ethnicity, cigarette usage, and sleep. In addition, 4 event types were defined from claims: diagnoses, procedures, provider specialty, and prescriptions. Further detail regarding demographics and events is found in eAppendix Figure 1. An indicator variable identified the year in which an event occurred.

1. **Lab results.** Results from 24 common lab tests (as identified by Logical Observation Identifiers Names and Codes number) were extracted for each year. Results were discretized in up to 7 ranges.
2. **Biometrics.** For each of the CMSS biometric screenings conducted, 6 variables were created (the 4 single-metric metabolic syndrome factors and systolic and diastolic blood pressure values). The values were then segregated into 7 ranges for blood pressure and 6 ranges for the remaining CMSS factors. In cases where the biometric corresponded to a lab test, the same discretization was used.
3. **Medication adherence.** We calculated a subject's medication possession ratio (MPR) for 4 classes of medication: antidiabetics, antihyperlipidemics, antihypertensives, and other cardiovas-

Figure 1. REFS Study Population Metabolic Syndrome Risk and Per Participant Per Year (PPPY) Medical Cost Profile



	0	1	2	3	4	5
Male, N	4730	5631	5170	3340	1655	554
Female, N	4562	4453	3591	2091	923	233
Allowed Cost PPPY—Male	3009	3785	4968	5758	6296	7273
Allowed Cost PPPY—Female	3930	4875	5448	6052	7317	6721

Metabolic syndrome was present in 26% of the 36,944 study subjects.

cular medications. More detailed information on MPR calculus is found in eAppendix Table 1. An MPR of 80% or higher was considered adherent.[4] For each year and each category of medication, a subject was categorized as: N/A (no prescriptions of that type), once and done (1 prescription of that type), not adherent, or adherent.

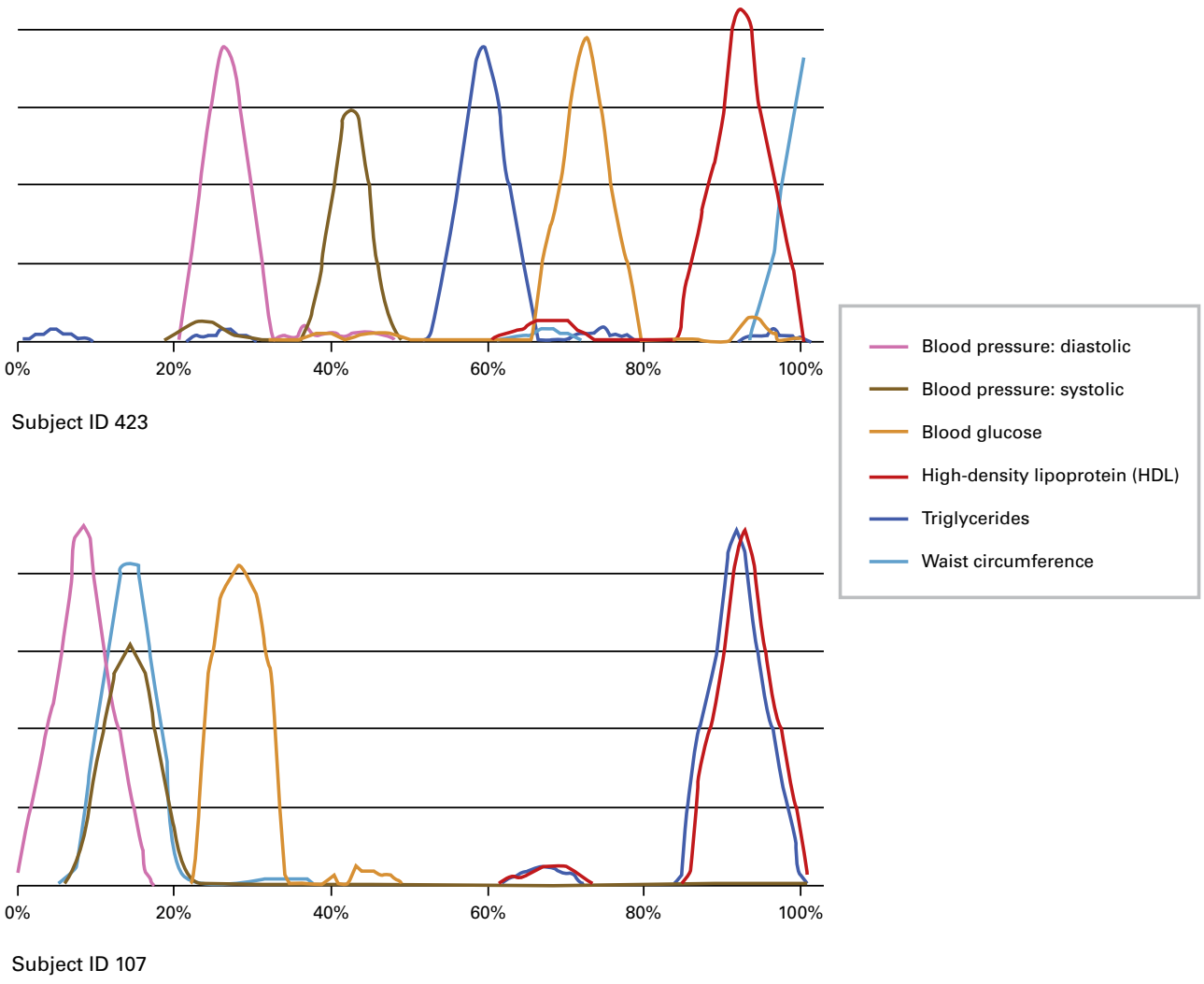
4. **Preventive visits.** A subject was deemed to have had a preventive visit if they had at least 1 claim during each year coded as a Preventive Visit (with one of 26 specific Evaluation & Measurement CPT-4 codes).

Statistical Methods: Platform Analytic Methods and Simulations

The REFS platform learns by Metropolis Monte Carlo[5] sampling from the posterior of the model-structure distribution. Model structure probabilities are computed in a Bayesian framework by marginalizing out the unknown parameter distributions against the observed data and maximum entropy parameter priors.[6] These model structure probabilities balance the model's fit of the data against the model's complexity.

Once learned, the model was interrogated by Forward Simulation (FS) to learn risk factors as well as the impact of interventions for individuals and populations. FS is a fast Monte Carlo process that samples simultaneously from the structure of the platform, the uncertainty in its parameters, and residual uncertainty on the outcomes that is efficient enough to be driven interactively. Multivariate categorical models were sampled describing each of the 6 discretized metabolic syndrome components. These models included up to 16 variables chosen from the total set of all variables. For each of the 6 metabolic syndrome components, the size of the space of models sampled during reverse engineering (RE) is the number of ways to choose up to 16 distinct variables from the 4291 variables possible or approximately 10^{44} models. Metropolis Monte Carlo can efficiently sample from these astronomically large hypothesis spaces guided only by the data even without prior knowledge to guide the search.

Figure 2. Subject Predicted Risk Factors Using Velocity Model



REFS velocity model produces individualized metabolic syndrome risk profiles.

Two models were learned. The Metabolic Syndrome Status Model was trained on claims-based events from 2010 to predict the CMSS measurements taken at the beginning of 2011 and the Metabolic Syndrome Velocity Model used claims-based events from 2011 *together with* the 2011 CMSS measurements to predict 2012 CMSS results.

Simulations Conducted

Because the number of model parameters is much larger than the number of observations, there were many models consistent with the observed data. The ensemble of models learned in the reverse engineering phase is a population sample from the posterior distribution over model structures.

Individual risk simulations. Forward simulations were computed for each of the 5 primary metabolic syndrome factors (with blood pressure separated into systolic and diastolic components for all study subjects) to predict likely values of metabolic syndrome factors at the next biometrics screening. The output for each factor was the probability of each range of the discretization of the factor. The probabilities across the outcome ranges were aggregated on either side of the factors out-of-range boundary and the resultant out-of-range probability computed for each factor. The individual out-of-range probabilities were further aggregated to compute the probability of metabolic syndrome.

Metabolic syndrome factor incremental perturbation simulations. To understand the impact of incremental changes in metabolic syndrome components on overall

Table 1. Impact of Incremental Changes in Each Metabolic Syndrome Factor

Metabolic Syndrome Factor	No. of Subjects Relevant	% of Total Subjects	Mean Change in Metabolic Syndrome Probability[a],%	Change Range, %	SD
Effect of Improvement					
Blood pressure	2282	6.2	–1.8	–13.9, .1	2.6%
Triglycerides	3276	8.9	–3.6	–25.1, 0	3.7%
Glucose	10,398	28.2	–5.6	–22.6, .1	3.9%
Waist circumference	16,490	44.6	–6.7	–40.7, .1	4.8%
HDL	4487	12.1	–5.4	–31.6, .1	5.4%
Effect of Incremental Progression (worsening)					
Blood pressure	2343	6.3	6.0	.1, 19.5	3.8%
Triglycerides	2635	7.1	8.0	.1, 37.9	5.3%
Glucose	15,096	40.9	6.9	0, 30.8	4.0%
Waist circumference	15,195	41.1	6.4	.1, 37.5	4.6%
HDL	1664	4.5	8.1	.2, 28.5	5.9%

HDL indicates high-density lipoprotein.
[a] *P* values for all of the "Mean Change in Metabolic Syndrome Probabilities" listed in Table 1 are less than 10.[6]

probability of metabolic syndrome for each individual, we simulated an incremental change in each metabolic syndrome component (a single range shift upward or downward for the component). From the 12 outputs, we recorded which incremental perturbation led to the largest increase and decrease in probability of metabolic syndrome for that patient, along with the magnitudes of the change in probability of metabolic syndrome.

Medication adherence and preventive visits simulations. The impact of medication adherence and preventive visits was assessed by counterfactual simulations of patients who were nonadherent in 1 or more of the drug-specific adherence metrics and patients who were noncompliant with preventive visits. For each patient the nonadherent metrics were switched to adherent and the patient-specific change in probability of metabolic syndrome was recorded. A similar simulation was applied to preventive visits.

RESULTS

Of the 2 models, the Status model was independent of CMSS measurements and relied solely on available demographic, medical claims, medication, and lab data to predict the outcome of the CMSS. In contrast, the Velocity model included the CMSS measurements in each year to predict the change in metabolic syndrome prevalence year over year. Both models predicted future risk of metabolic syndrome on both a population level and an individual level, both in aggregate and by specific metabolic syndrome risk factor, with good to excellent predictive ability. The predictability was slightly higher with the Velocity model; overall, receiver operating characteristic/area under the curve (ROC/AUC) = 0.80 for the Status model and 0.88 for the Velocity model (supporting detail in eAppendix Tables 2 and 3).

The ability of the models to produce highly individualized risk profiles for overall risk of metabolic syndrome and by specific risk factors allows for more successful patient engagement in subsequent care management programs. Figure 2 shows 2 different individual risk profiles. Subject ID 423262 was a 46-year-old male with current out-of-range metabolic syndrome risk factors of high-density lipoprotein (HDL) and waist circumference. He had a 92% predicted probability of developing metabolic syndrome within 12 months, and a 73% probability of developing abnormal blood glucose as a third specific metabolic syndrome risk factor during the study period. Subject ID 107975 presented a contrasting profile. He was a 37-year-old male with 2 out-of-range metabolic syndrome risk factors—HDL and triglycerides—but had only a 40% predicted probability of developing metabolic syndrome within 12 months. For this subject, abnormal blood glucose was also the most likely abnormal factor to develop next, but carried only a 26% likelihood.

Looking at the modeled effect of incremental changes in individual out-of-range metabolic syndrome risk factors on the subsequent risk of developing metabolic syndrome within 12 months, the factors with the largest weighted effect were waist circumference and glucose (Table 1). A similar pattern was seen when looking at the effect of

Table 2. Impact on Cost to Treat of 1% Change in Each Metabolic Syndrome Factor

Metabolic Syndrome Factor	PPPY Change in Medical Cost, $
Blood glucose	27.08
Waist circumference	13.78
Blood pressure	7.91
Triglycerides	0.64
HDL	0.39

HDL indicates high-density lipoprotein; PPPY, per patient per year.

incremental (1%) improvement in a given metabolic syndrome risk factor on subsequent healthcare costs (Table 2).

We also modeled the effect of 2 separate surrogates of medically adherent behavior on the subsequent risk of developing metabolic syndrome: adherence with specific prescribed medications detailed above, and adherence with routine scheduled preventive doctor visits (Table 3). On a population basis, a clear benefit was derived from improved adherence to preventive visits; 87% of previously nonadherent individuals showed a modest decrease of up to 10% in subsequent metabolic syndrome risk.

DISCUSSION

The application of advanced analytics to large data sets is becoming more prevalent in healthcare as the volume and variety of data produced expand and the cost of sophisticated analytic solutions drops. McAna et al reported that generating predictive models using generally available administrative data and statistical software like Stata could accurately calculate hospitalization risk in populations of Medicaid enrollees to support interventions by care managers.[7] Similarly, the analytic platform discussed here allows for personalized risk predictions and the rapid development of data-driven, targeted interventions for individuals with or at risk of metabolic syndrome which can help improve population and individual health, and reduce costs.[8]

The platform provides healthcare researchers and managers with an additional option and unique tool to generate insights faster and increase program impact and returns. The platform can learn models directly from data which capture the underlying mechanisms and processes consistent with the data. In addition, the platform's underlying methodologies are data agnostic and allow for extreme data heterogeneity (including missing data). These characteristics allow the platform to yield more naturally interpretable answers (in terms of probabilities) and more realistic predictions (by incorporating uncertainty). Additionally, by integrating over model parameters, the platform safeguards against over-fitting. Finally, the platform is unbiased as to sample size, and is equally applicable to minimal samples and more complicated models which traditional approaches are unable to estimate.

These distinctive features benefit researchers by allowing them to:

- Generate insights faster, because healthcare organizations can immediately use the platform on available data and extract and validate actionable insights. The insights derived in the study outlined here were reached in 3 months (1.5 months per model), as opposed to the years that clinical trial and longitudinal studies take.
- Improve intervention program design, impact, and returns. The speed at which researchers can simulate counterfactual scenarios with the platform allows them to assess the potential impact of specific interventions before investing in comprehensive programs, to create individualized targeting based on personalized data, and to build intervention models that dynamically learn where to adjust programs.

Other models differ in critical aspects. As an example, and in contrast, Archimedes models capture the current state of clinical knowledge as a set of algorithmic processes and are thus not responsive to customer data and are limited to what is already known. General purpose statistical analysis platforms such as SAS can generate learning models from data through stepwise regression, but also require expert oversight and guidance.

Our results confirm earlier studies[9,10] that identified reduction in waist circumference (or weight) as the primary factor in decreasing both the risk of developing metabolic syndrome and future costs, and also confirmed that improved adherence to prescribed medication for control of blood pressure, lipids, and diabetes mellitus will reduce the subsequent risk of developing metabolic syndrome. Finally, the results of improved adherence to routine, scheduled preventive visits demonstrated a modest but clear benefit as well: individuals improving adherence achieved less than a 10% decrease in risk, yet almost 90% of individuals with improved adherence to preventive visits showed some decrease in metabolic syndrome risk.

As a result of this analysis, Aetna Clinical Innovation Labs is launching a novel metabolic syndrome intervention pilot specifically focusing on reducing waist circumference. As primary end points, the year-long pilot will measure weight loss and reduction in metabolic risk, both overall and by risk factor. Secondary end points include medical utilization and cost metrics.

■ **Table 3.** Impact of Adherence to Select Medications and Preventive Visits on Subsequent Metabolic Syndrome Risk

	Effect of Adherence	Total Study Population	% of Total
Impact of medication adherence on subsequent metabolic syndrome risk	Positive effect	3304	58.25
	Negative effect	2368	41.75
Impact of preventive visit on subsequent metabolic syndrome risk	Positive effect	18,746	87
	Negative effect	2857	13
Medication Adherence			
Mean Delta Metabolic Syndrome	–0.7%		
SE	0.207%		
P	.00024		
Preventive Visit Adherence			
Mean Delta Metabolic Syndrome	–2.1%		
SE	0.018%		
P	<10-16		

Potential Limitations of this Study

Results and conclusions are derived from the data of a single large nationwide employer with, presumably, unique socioeconomic, demographic, and clinical characteristics. Therefore, the results should not be directly extrapolated to other populations. Furthermore, although the results presented appear to be predictive of future metabolic syndrome risk, this is based on retrospective analysis of historical data. We would encourage that this general analytic approach and the various associated findings be validated by prospective analyses in multiple and varied patient populations.

A prospective validation of the hypothesis from the pilot referenced above is also needed. And, although the platform can be used to derive causality inferences, this was not done in this study; accordingly, the results presented should be regarded as associative and hypothesis-generating. This study was not meant to take the place of gold standard, clinical trials, and was not initiated or approved by an institutional review board.

Because there are multiple big data analytic techniques and platforms available, and it is unclear which could better answer specific clinical questions, a direct head-to-head comparison of various analytic methods using standardized data sets and a priori agreed-upon metrics would help to answer this important question.

CONCLUSIONS

We applied a proprietary "big data" analytic platform to a large healthcare data set of a single nationwide employer to test predictive models relative to metabolic syndrome. The models were generated within 3 months, and predicted the subsequent 12-month risk of metabolic syndrome at both a population and individual level, and by overall metabolic syndrome risk as well as by individual metabolic syndrome risk factor. The rapid availability of accurate, individualized models is being used to develop personalized clinical outreach and engagement strategies for affected individuals, which we believe will decrease the clinical burden of metabolic syndrome and its associated costs.

We believe this study demonstrates how big data analytic techniques can be applied to large complex data sets to generate pragmatic, actionable insights relative to metabolic syndrome and other clinical conditions.

Acknowledgments

The authors thank the management team at the retailer referenced in this study for insights related to its employee population as well as employee care management strategies. We are thankful for the insight and guidance from Michael Palmer, head of innovation at Aetna. We are appreciative of the contributions from our colleagues at GNS Healthcare as well, including John Kucera, Casey Marks, and Karl Runge. We also are grateful for support received from Accenture's Health & Public Services practice, specifically Tom Heatherington, Lokesh Gurunathan, and Kerry Vincent, each of whom provided additional insights regarding big data applications. Finally, we appreciate the help of KC Spears of Compel Communications for her editorial and submission support.

Author Affiliations: Aetna Innovation Labs, Hartford, CT (GBS, ABS); GNS Healthcare, Cambridge, MA (BWC, CJM); Accenture Health and Public Services, Minneapolis, MN (BPK).

Source of Funding: Funding for development of this article was provided by GNS Healthcare and Aetna Innovation Labs.

Author Disclosures: Mr Steinberg and Mr Scott report employment with Aetna, who helped fund this study. Dr Church and Ms McCall report employment with GNS Healthcare, who also helped fund this study. Mr Kalis reports employment with Accenture Health, which was hired on a consulting basis to support the research associated with this study. Views expressed in this article about the applicability of REFS to a broad

array of data sets and biodmedical information are solely those of the authors, and do not represent the official view of Accenture.

Authorship Information: Concept and design (GBS, BWC, CJM, ABS); acquisition of data (ABS); analysis and interpretation of data (GBS, BWC, CJM, BPK); drafting of the manuscript (GBS, BWC, CJM, ABS, BPK); critical revision of the manuscript for important intellectual content (CJM, ABS); statistical analysis (CJM); obtaining funding (ABS); administrative, technical, or logistic support (CJM, BPK); supervision (GBS, CJM).

Address correspondence to: Gregory B. Steinberg, MB, BCh, Aetna Innovation Labs, PO Box 359, Dingmans Ferry, PA 18328. E-mail: gsteinberg@aetna.com. Bruce W. Church, GNS Healthcare, 1 Charles Park, Third Floor, Cambridge, MA 02141. E-mail: bruce@gnshealthcare.com.

REFERENCES

1. Fitch K, Pyenson B, Iwasaki K. Metabolic syndrome and employer sponsored medical benefits: an actuarial analysis. Milliman website. http://publications.milliman.com/research/health-rr/pdfs/metabolic-syndrome-employer-sponsored-RR03-01-06.pdf. Published March 2006.

2. Ervin RB. Prevalence of metabolic syndrome among adults 20 years of age and over, by sex, age, race and ethnicity, and body mass index: United States, 2003-2006. National Health Statistics Report, vol 13. CDC website. http://www.cdc.gov/nchs/data/nhsr/nhsr013.pdf. Published May 5, 2009.

3. Edington DW, Schulz AB. Metabolic syndrome in a workplace: prevalence, co-morbidities, and economic impact. *Metab Syndr Relat Disord.* 2009;7(5):459-468.

4. Martin BB, Wiley-Exley E, Richards S, Domino M, Carey T, Sleath B. Contrasting measures of adherence with simple drug use, medication switching, and therapeutic duplication. *Ann Pharmacother.* 2009; 43(1):36-44.

5. Metropolis N, Rosenbluth A, Rosenbluth M, Teller A, Teller E. Equation of state calculations by fast computing machines. *J of Chemical Physics.* 1953;21(6):1087-1092. http://www.aliquote.org/pub/metropolis-et-al-1953.pdf.

6. Jaynes ET. *Probability Theory: The Logic of Science.* New York, NY: Cambridge University Press; 2003.

7. McAna J, Crawford A, Novinger B, et al. A predictive model of hospitalization risk among disabled Medicaid enrollees. *Am J Manag Care.* 2013;19(5):e166-e174.

8. Target big data to gain access to future strategic leaders. Gartner Market Analysis website. http://www.gartner.com/id=2219715. Published October 31, 2012.

9. Heidari Z, Hosseinpanah F, Mehrabi Y, Safarkhani M, Azizi F. Predictive power of the components of metabolic syndrome in its development: a 6.5-year follow-up in the Tehran Lipid and Glucose Study (TLGS). *Eur J Clin Nutr.* 2010;64(10):1207-1214.

10. Goodpaster BH, DeLany JP, Otto A, Kuller L et al. The effects of aerobic, resistance, and combined exercise on metabolic control, inflammatory markers, adipocytokines, and muscle insulin signaling in patients with type 2 diabetes. *Metabolism.* 2011;60(9):1244-1252. ■

www.ajmc.com Published as a Web Exclusive