Editor: **Daniel B. Neill,** H.J. Heinz III College, Carnegie Mellon University, neill@cs.cmu.edu

Information Visualization for Chronic Disease Risk Assessment

Christopher A. Harle, *University of Florida*
Daniel B. Neill and Rema Padman, *Carnegie Mellon University*

According to the US-based Institute of Medicine (IOM) and the National Research Council (NRC), the vision for 21st century healthcare includes increased attention to "cognitive support, which refers to computer-based tools and systems that offer clinicians and patients assistance for thinking about and solving problems."[1] Tools for cognitively guided decision support are particularly important for primary care clinicians who manage patients with complex chronic diseases, such as diabetes. Chronic disease prevalence, complication rates, and costs are alarmingly high, with many people affected by multiple conditions and disabilities.[2] Furthermore, primary care clinicians face severe time constraints when providing care,[3] and their reimbursement is increasingly tied to quality of care and patient outcomes.[4]

Information visualization—"the use of computer-supported, interactive visual representations of abstract data to amplify cognition"[5]—has supported many tasks in chronic disease care. Visualizations include color-coded spreadsheet displays,[6] line or bar graphs that show concordance between individual patient measures and benchmark values,[7,8] bar graphs that summarize quality measures across many patients,[7] clinical event timelines,[8,9] and combinations of these approaches. These tools typically display multiple attributes of a single patient or summarize many patients on a single dimension. However, existing tools are limited in their ability to analyze and display high-dimensional data for many patients in a single view.

Here, we examine a new approach for quantitatively summarizing and visually displaying information on many relevant health dimensions across many patients. Specifically, we describe a method and prototype software tool that provide two-dimensional visualization and risk classification across multiple risk factors. This framework lets clinicians interactively visualize patient information at the population, subpopulation, and individual patient levels. We quantitatively evaluate our framework's ability to visually classify a patient population according to risk of heart attack, thus expanding on its previous application in the context of diabetes onset risk.[10]

Materials and Method Overview

We developed our methodology and prototype software tool for visualizing and stratifying high-dimensional patient data using a combination of statistical machine learning and information-visualization techniques. We developed the prototype's visualization interface in Adobe Flex and intend to implement it in JavaScript/HTML5 for improved scalability and customization. All computations are executed by R statistical software libraries, using Rserve[11] to communicate with the visualization application's Java back end. We tested our tools using an anonymous dataset from the American Diabetes Association (ADA). This dataset contained $M = 22$ demographic and clinical variables related to heart attack risk for $N = 588$ people with type 2 diabetes. The variables included age, sex, ethnicity, smoking status, body mass index (BMI), systolic blood pressure (SBP), diastolic blood pressure (DBP), LDL cholesterol, HDL cholesterol, most recent HbA1c value, physical activity level, whether the person sees a physician regularly, medical history of angina, albuminuria, cardiac bypass, heart attack, congestive heart failure, and heart disease, and medications currently taken (such as aspirin, antihyperglycemics, antihypertensives, or antihyperlipidemics). The data also included each person's risk of having a

1541-1672/12/$31.00 © 2012 IEEE
Published by the IEEE Computer Society

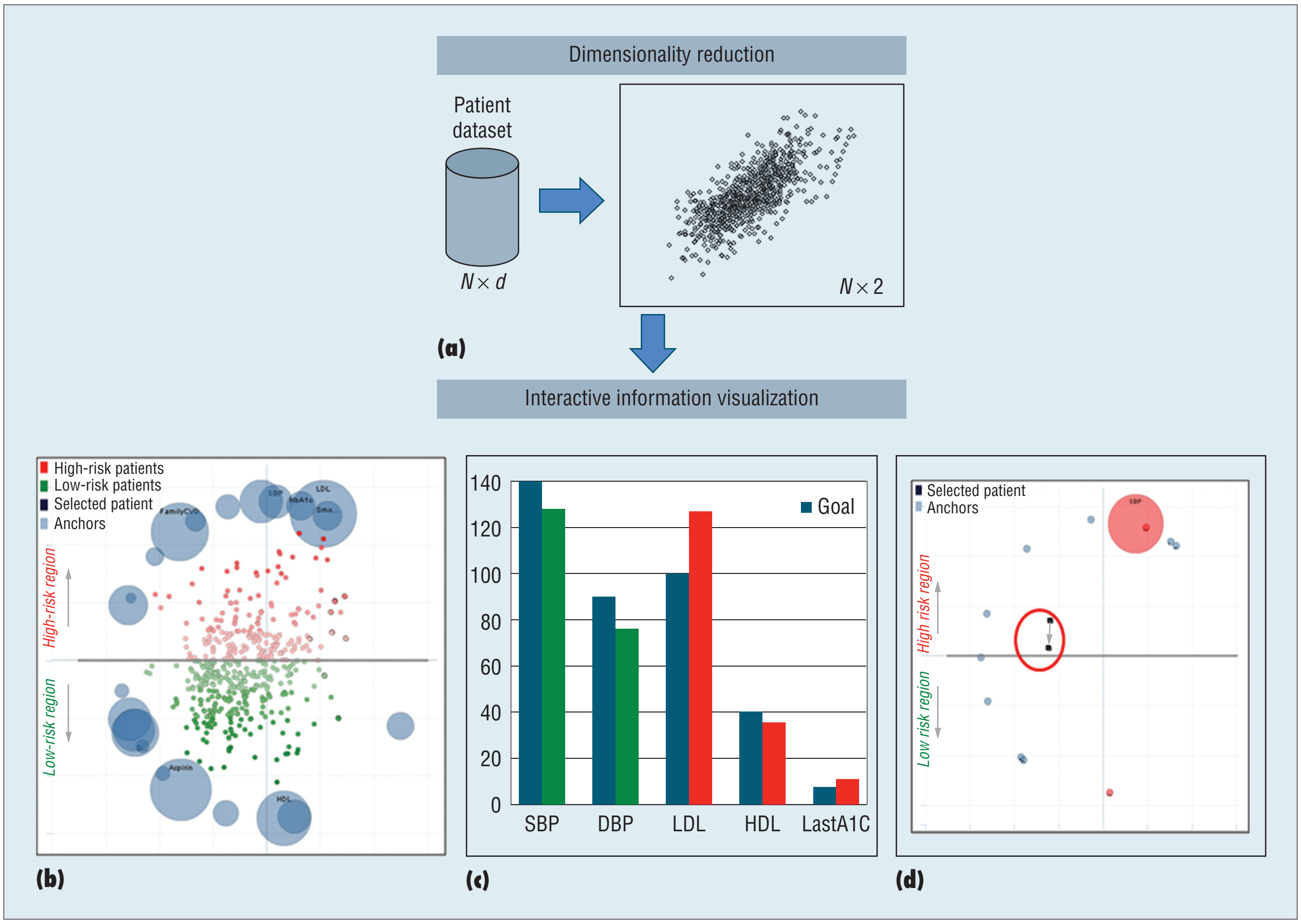

Figure 1. Visual risk assessment and analysis process. Our approach (a) projects patient data into two dimensions for plotting. Clinicians can (b) view the patient population risk, (c) assess an individual patient's risk factors vs. goal values, and (d) visualize the effect of a hypothetical intervention on an individual patient's risk.

heart attack at some point in the next 30 years, estimated using the Archimedes model.[12] However, the risk classification approach we describe here is not limited to 30-year risk nor does it rely on the Archimedes model. It can use any measure of risk for which labeled data is available.

Methodology and System Development

First, we projected and plotted the data in two dimensions using two statistical dimensionality reduction techniques, *principal component analysis* (PCA) and *linear discriminant analysis* (LDA) (Figure 1a). We used PCA, plotted on the x-axis, because data transformed using the first principal component would be maximally scattered in the resulting one-dimensional space.[13] This visually highlights clinical differences between patients in a view that is simpler than comparing patient data across many variables, such as in a large spreadsheet. LDA, on the other hand, is explicitly concerned with classification and thus is useful for stratifying patients by risk. Given labeled training data from two distinct classes (for instance, *low risk* of heart attack versus *high risk* of heart attack), LDA computes a linear data transformation that maximizes the ratio of between-class variance to within-class variance. We used LDA, plotted on the y-axis, because high- and low-risk patients should be well-separated visually when clinicians view the data.[14] The first principal component and linear discriminant together define a linear transformation of each high-dimensional data point Z_j into the 2D visualization space: $(x_j, y_j)^T = DZ_j$, where D is a $2 \times M$ matrix with the first row determined by the first principal component and the second row determined by the linear discriminant; Z_j is a column vector of length M, and x_j and y_j are scalars representing the new coordinates in 2D space.

We next applied information-visualization techniques to visually recapture some of the information lost by dimensionality reduction and support cognitively guided decision making. First, we added "attracting anchors"[15] to the plot to convey each risk variable's coefficient weights $D_{1,i}$

and $D_{2,i}$ on the principal component and the linear discriminant, respectively (Figure 1b), indicating each variable's impact on risk and variation at the population level. The anchors are represented by larger circles on the perimeter of the patient data point cloud. In terms of polar coordinates, we determine each variable's anchor location by the angle, θ_i, given as follows:[10]

$$\theta_i = \begin{cases} \tan^{-1}\left(D_{2,i} / D_{1,i}\right) & D_{1,i} \geq 0 \\ \pi + \tan^{-1}\left(D_{2,i} / D_{1,i}\right) & D_{1,i} < 0. \end{cases} \tag{1}$$

Each anchor's location conveys the relative size of the two components' coefficients for each variable, which we call the *direction of attraction*. We express the relative size of the component coefficients between risk factor variables, or the *magnitude of attraction*, using each anchor's area. We define the anchor area S_i to be proportional to the length of the vector formed by the weighting coefficients $D_{1,i}$ and $D_{2,i}$, given as follows:[10]

$$S_i = C\left(\sqrt{D_{1,i}^2 + D_{2,i}^2}\right),$$

where S_i is the area of the plotted anchor and C is a tuning constant set such that the smallest anchors are large enough to be visible. We can use the anchors' size and direction to interpret which risk factor variables have relatively large or small coefficients on each component. For example, if variable i has a large positive coefficient on the first principal component and a large positive coefficient on the linear discriminant, it will be represented by a relatively large anchor located in the plot's upper-right quadrant. Patient data points with larger values for variable i will then be plotted further toward the plot's upper-right quadrant. Our rationale for the anchors was to convey in a single view an understanding of which variables are most influential on risk at the population level and which individuals tend to have higher values for individual risk factor variables.

Finally, we added interactive information-visualization techniques to the prototype tool to enable zooming, filtering, and viewing individual patient details as recommended for exploratory information-visualization tasks[16] (Figures 1b–d). In a clinical context, these techniques let users focus their attention on and analyze smaller subpopulations or individual patients as needed, based on their disease risk or other clinical characteristics. When users select an individual patient, they can compare modifiable risk factors' current values versus goal values for that patient (Figure 1c). The software resizes the anchors to visually depict each risk factor's relative impact for that particular individual (Figure 1d). The system also includes a hypothetical intervention feature that lets users see how prospective interventions or changes in risk factor values would impact the patient's location on the population risk graph (Figure 1d).

Evaluation

Our quantitative evaluation of our framework focused on its utility for stratifying patients into risk groups. Based on the 30-year heart attack risk probabilities contained in the sample data, we labeled half the data low risk and half high risk of future heart attack. We used naïve Bayes, logistic regression, k-nearest neighbor (k-NN), and support vector machine (SVM) approaches to classify the data. To understand how well our method visually separated low- and high-risk patients, we compared the classification accuracies of these other common methods to that of applying LDA to the data plotted in two dimensions using our described approach. We calculated the classification error for each method using 10-fold cross-validation.

Figure 2 shows the visual classifier and includes an LDA-based decision boundary, which separates the patients predicted to be high risk and those predicted to be low risk. The 10-fold cross validation error from the visual classifier was .13, which was similar to or less than the error found after applying other commonly used classification methods to the full-dimensional dataset, including naïve Bayes (.22), logistic regression (.11), 1-NN (.21), 10-NN (.20), 200-NN (.29), and SVM (.12).

Our framework presents a new approach to organizing and delivering "cognitively guided" data to clinicians. The methodology combines simple yet powerful statistical methods with intuitive data visualizations. The visual classifiers provide a contextualized and potentially more interpretable means of communicating risk information on complex patient populations to time-constrained clinicians. When using the interactive software tool described here, users can navigate from a 2D patient population view to views of individual patients to quickly assess their risk and risk factors, and the effects of interventions.

Our evaluation used a quantitative analysis of classification accuracy and our interpretations of clinical insights depicted in the heart attack risk models. Subsequent interviews and demonstrations with a small sample of diabetes educators, physicians, and patients also suggested these tools' potential usefulness for personalized clinical interventions, risk communication, and patient education. These are necessary first steps in evaluation, but formal clinical usability and usefulness testing is warranted. Through user testing, we might further revise our methods to better align with clinician information needs and decision-making processes.

We must also evaluate our visualization models' classification accuracy using other data, including data that comes directly from clinical sources. To achieve this, we plan to acquire longitudinal data containing both risk factors and patient outcomes. Moreover, prior to applying visual classifiers to any patient population that differs significantly from those on which we originally validated them, our models would need additional training and calibration, as has been recommended with other risk assessment models. Finally, this visualization method and tool could extend to other risk-assessment scenarios in healthcare delivery, such as evaluating risks to patient safety and clinical guideline compliance.

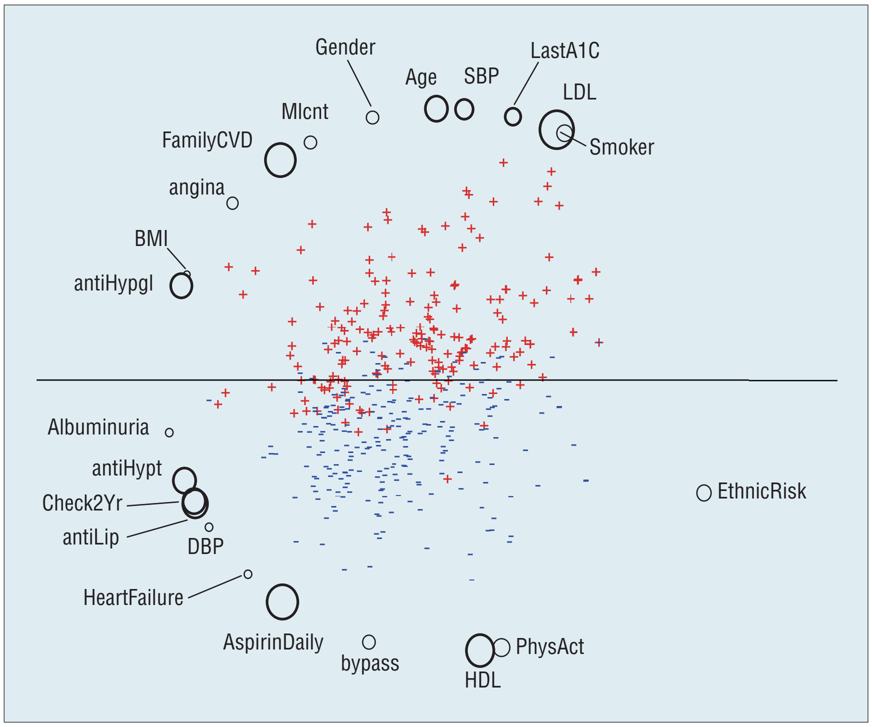

Figure 2. Heart attack risk visualization showing classification accuracy on training data. The *x*-axis shows the first principal component, and the *y*-axis shows the linear discriminant. We plotted predicted high- and low-risk patients above and below the horizontal boundary, respectively. We colored labeled high- and low-risk patients red and blue, respectively. Twenty-two risk factor variables relevant to prediction and decision making are anchored around the patient data points.

Acknowledgments

We thank the American Diabetes Association for sharing data; Michelina Fato, David Eibling, Francis Solano of the University of Pittsburgh Medical Center (UPMC) and Janice Zgibor and Ken Sochats of the University of Pittsburgh for their input on the visualization methods and prototype system; Ambarish Joshi for his work on the prototype implementation; and the UPMC Technology Development Center for partially funding our implementation initiative.

References

1. *Computational Technology for Effective Healthcare: Immediate Steps and Strategic Directions*, Committee on Engaging the Computer Science Research Community in Health Care Informatics, Nat'l Research Council, W.W. Stead and H.S. Lin, eds., National Academies Press, 2009.
2. "2011 National Diabetes Fact Sheet," US Centers for Disease Control and Prevention, 2011; www.cdc.gov/diabetes/pubs/pdf/ndfs_2011.pdf.
3. T. Ostbye et al., "Is There Time for Management of Patients with Chronic Disease in Primary Care?" *Annals of Family Medicine*, vol. 3, no. 3, 2005, pp. 209–214.
4. M.B. Rosenthal et al., "Climbing Up the Pay-for-Performance Learning Curve: Where Are the Early Adopters Now?" *Health Affairs*, vol. 26, no. 6, 2007, pp. 1674–1682.
5. S.K. Card, J.D. Mackinlay, and B. Shneiderman, *Information Visualization: Using Visualization to Think*, Morgan Kaufmann, 1999.
6. K. Keshavjee et al., "Compete III: A Chronic Disease Management Program for Diabetes and Vascular Disease," Ann. Symp. Am. Medical Informatics Assoc., featured presentation, 2006; www.compete-study.com/documents/COMPETE-III_CDSS_Theatre_Style_Demo_Nov_13_2006.pdf.
7. J.A. Linder et al., "Improving Care for Acute and Chronic Problems with Smart Forms and Quality Dashboards," *Proc. 2006 Ann. Symp. Am. Medical Informatics Assoc.*, Am. Medical Informatics Assoc., 2006, p. 1193.
8. C. Plaisant et al., "Searching Electronic Health Records for Temporal Patterns in Patient Histories: A Case Study with Microsoft Amalga," *Proc. 2008 Ann. Symp. Am. Medical Informatics Assoc.*, Am. Medical Informatics Assoc., 2008, pp. 601–605.
9. T.D. Wang et al., "Visual Information Seeking in Multiple Electronic Health Records: Design Recommendations and a Process Model," *Proc. ACM Int'l Health Informatics Symp.*, ACM, 2010, pp. 46–55.

10. C.A. Harle, D.B. Neill, and R. Padman, "An Information Visualization Approach to Classification and Assessment of Diabetes Risk in Primary Care," *Proc. 3rd INFORMS Workshop on Data Mining and Health Informatics* (DM-HI 08), J. Li, D. Aleman, R. Sikora, eds., INFORMS, 2008; www.cs.cmu.edu/~neill/papers/dmhi08.pdf.
11. S. Urbanek, "Rserve: A Fast Way to Provide R Functionality to Applications," *Proc. 3rd Int'l Workshop Distributed Statistical Computing*, Technische Universität Wien, 2003; www.ci.tuwien.ac.at/Conferences/DSC-2003/Proceedings/Urbanek.pdf.
12. D.M. Eddy and L. Schlessinger, "Archimedes: A Trial-Validated Model of Diabetes," *Diabetes Care*, vol. 26, no. 11, 2003, pp. 3093–3101.
13. I.T. Jolliffe, *Principal Component Analysis*, 2nd ed., Springer, 2002.
14. R.O. Duda, P.E. Hart, and D.G. Stork, *Pattern Classification*, 2nd ed., John Wiley & Sons, 2001.
15. K.A. Olsen et al., "Visualization of a Document Collection: The VIBE System," *Information Processing and Management*, vol. 29, no. 1, 1993, pp. 69–81.
16. B. Shneiderman, "The Eyes Have It: A Task by Data Type Taxonomy for Information Visualizations," *Proc. IEEE Symp. Visual Languages*, IEEE, 1996, pp. 336–343.

Christopher A. Harle is an assistant professor of health services research, management, and policy at the University of Florida's College of Public Health and Health Professions. Contact him at charle@phhp.ufl.edu.

Daniel B. Neill is the Heinz Career Development chair and associate professor of information systems at Carnegie Mellon University's Heinz College, where he directs the Event and Pattern Detection Laboratory. Contact him at neill@cs.cmu.edu.

Rema Padman is a professor of management science and healthcare informatics and a Thrust Leader of Health IT Research at iLab in Carnegie Mellon University's Heinz College. Contact her at rpadman@cmu.edu.