Resolving Public Expert Model Disagreement in Medical Risk Evaluation

Anthony (Tony) J Grichnik	Dr. Michael L. Taylor	Dr. Christos Nikolopoulos	James R. Mason
Caterpillar Inc.	Caterpillar Inc.	Bradley University	eServ
grichaj@cat.com	Taylor_Mike_L@cat.com	chris@bumail.bradley.edu	Mason_James_X@cat.com

Abstract

Medical Risk Stratification (MRS) models capture the relationships between modifiable and unmodifiable factors to help us understand how various risk factors jointly contribute to the likelihood of contracting a disease in the future. While multiple MRS models exist for diseases like cardiovascular disease (CVD) and diabetes, these models often conflict with one another when applied to real-world populations outside their original study groups. In this paper we examine the conflict between public MRS models of CVD and diabetes and quantify the disagreement using both simulated and real-world populations. A process to resolve these conflicts is presented and the resulting improvement in predictive power is quantified through Bayesian Posterior Probability (BPP) analysis of a population prior to disease onset. By producing improved MRS models, we provide valuable knowledge to those charged with improving the health of populations under their care.

Keywords: Medical risk stratification, neural networks, genetic algorithms, public risk model disagreement, Bayesian posterior probability.

1. Introduction

A large body of work exists to support the medical risk stratification (MRS) of individuals in a population. These MRS models cover diverse issues such as diabetes, cardiovascular disease, stroke, various kinds of cancers, depression, suicide, and eating disorders - to name only a few. Health promotion programs sponsored by corporations and public entities are beginning to leverage these models on a large scale to assess the overall health of their respective populations. This is made possible by low-cost computing platforms of increasing power and by the increasing depth of historical data in data warehouses available to health promotion programs. Corporations are studying results from these models to gain insight into ways to improve the health of their workforce. Along the way, the initiatives of these individual corporate and public entities also have a broad, positive impact on the community. Corporately sponsored programs for active employees affect retired employees as well as the dependents and spouses of both employees and retirees. As an example, for every

L. Magdalena, M. Ojeda-Aciego, J.L. Verdegay (eds): Proceedings of IPMU'08, pp. 568–575
Torremolinos (Málaga), June 22–27, 2008

active employee participating in the Caterpillar® Healthy Balance® health promotion program in the Peoria area, there are about four additional persons engaged by the Healthy Balance program that are not on the current payroll. When combined, this equates to about 20% of the general Peoria-area population.

Broadly, MRS models are based on two types of data.

- Single factor studies – A large number of studies have been done to measure the influence of a single factor on chronic diseases. One recent example studied the influence of post-menopausal hormones on cardiovascular disease (CVD) and found an important relationship that ended a long-running trial when significant risks were discovered [1]. One challenge of such studies is to isolate the effect of uncontrolled factors on the experimental results. As a result, relatively little knowledge is gained about the interactions among factors.

- Population derivative studies – Many famous studies on chronic diseases examine a broad range of factors over time in well-bounded populations. For instance, the Framingham Heart Study continues to track the history of the residents of Framingham, Massachusetts, USA and their incidence rate of CVD over a period of more than 50 years [2]. By studying a larger population, many factors can be extracted and examined, including the discovery of interactions among factors. One disadvantage of these studies is that they often reflect the population bias in the sample. In the Framingham case, the study population has relatively few non-Caucasian persons and few female patients. This limits the conclusions that can be drawn from such a population.

2. Study Data and Information Privacy

Two diseases of interest to the Healthy Balance program's population are CVD and diabetes. Several public models exist to stratify the risk for both diseases. Using information from self-survey forms provided by participants, medical claims history, prescription drug benefit payments and general demographic information we can extract the specific input variables required by public MRS models for many individual cases. Collectively these produce the population's risk profile.

One challenge in this process is maintaining data privacy. Both Caterpillar in general and the Healthy Balance program in particular have strict data privacy requirements [3]. To comply with these policies, we enforce the following processes.

- All data streams have personally identifiable information removed. Such information is replaced by a unique code to allow alignment across data systems. This unique code is not used in any other system at Caterpillar.

- Measurements are reported in aggregate form only. This ensures that individual records – which are already made anonymous – are further obscured.

- Aggregates are not reported if the sample size drops below 50. This prevents very small aggregates from becoming personally identifiable.

- All access to the data requires authentication and is logged regularly.

3. Measuring Disagreement in Public MRS Models

MRS models for the same disease or condition can disagree in the following ways.

Variable list disagreement – Different factors may have been considered between two or more models of the same disease. Variable list disagreement can be detected easily by inspection

Variable sensitivity disagreement – If two or more models use the same variable as input, they may disagree on its risk contribution. Knowledge of the underlying mathematics of the model is required to detect this type of disagreement.

Variable interaction disagreement – If two or more models have two or more variables in common, the interaction may affect the risk stratification sensitivity in different ways. This type of disagreement can be extremely difficult to detect, especially in complex models.

Variable substitution disagreement - As with any large, real-world data stream there will be missing values. When missing values occur, we substitute a condition that will produce a lowered response in the risk score[1]. The same substitution is used for both models whenever the variable lists overlap. In cases where the substitution affects one model but not the other, we also substitute a value that produces a lower score. This disagreement cannot be avoided in an experiment using real-world data, but it can be avoided when using simulated data.

Table 1 catalogs the pairs of models studied to demonstrate these types of disagreement. Variable list disagreement is illustrated in Tables 2 and 3. The public MRS models do not always use the same scoring system. For instance, the HSPH CVD model ranks risk on a scale of -65 to 130 for our test population while the AHA/NIH CVD model of the same disease ranks risk on a scale of 0 to 50 for that same population. We evaluate results on a full-scale reading percentage (FSR%) basis, in which the minimum risk is assigned a score of 0% and the maximum risk is assigned a score of 100%. Since each member of the pair of models for each disease are ostensibly models of the same disease, we expect that the resulting risk scores would be highly correlated on an FSR% basis.

In our first set of experiments, we drove the pairs of public risk models with simulated population data. In this case, we eliminate the bias from variable substitution disagreement by ensuring that the simulated population has no missing

[1] We choose to bias our variable substitutions to produce lowered scores because of the "first do no harm" philosophy in medicine. This may seem counterintuitive at first. Someone who believes they are at high risk may engage in medical tests or procedures that are unwarranted based on the known data. These tests and procedures have real, measurable risks. Overall, our goal is to recommend action when the evidence supports it and to withhold our action recommendation when the evidence is unclear or missing.

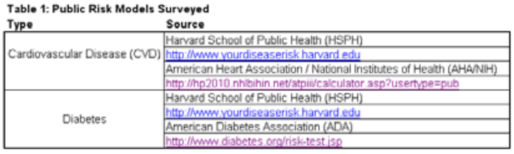
values[2]. The results of our simulations are shown in Figures 1 and 2. As these figures illustrate, our simulated populations generate results with significant disagreement.

In our second set of experiments, we drove the same pairs of models with a sample of real-world data from the Healthy Balance program's population. This random sample used approximately 10% of the available data in our data warehouses. Persons already diagnosed as positive for either disease were ignored and are not included in our results for that model. Variable substitution rules described previously were applied. The resulting disagreement is shown in Figures 3 and 4. These results reinforce our observations from the simulated population.

Finally to eliminate any possible effect of sample bias in the aforementioned results with the real-world population data, we applied all of the case data in the data warehouse to the respective public risk models while observing the rules described in the second set of experiments. The resulting disagreement is shown in Figures 5 and 6.

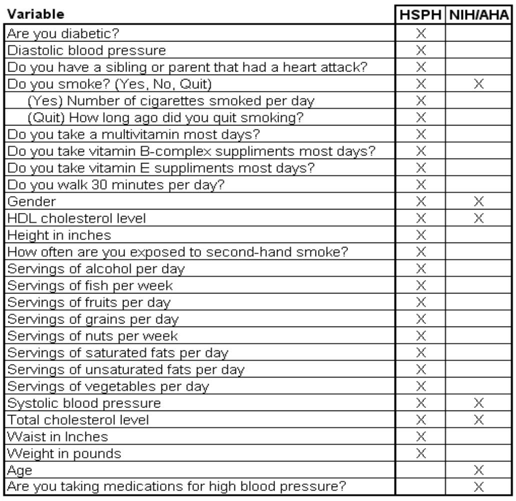
In all three experiments – simulated populations, sampled real-world populations, and complete real-world populations – we discover significant differences in the risk stratification of CVD and diabetes among public MRS models. A summary of these differences is contained in Table 4.

Table 1: Public Risk Models Surveyed



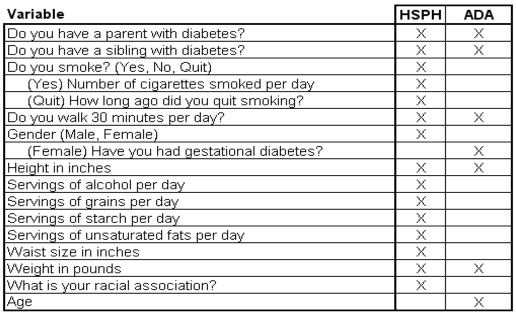
Type	Source
Cardiovascular Disease (CVD)	Harvard School of Public Health (HSPH)
	http://www.yourdiseaserisk.harvard.edu
	American Heart Association / National Institutes of Health (AHA/NIH)
	http://hp2010.nhlbihin.net/atpiii/calculator.asp?usertype=pub
Diabetes	Harvard School of Public Health (HSPH)
	http://www.yourdiseaserisk.harvard.edu
	American Diabetes Association (ADA)
	http://www.diabetes.org/risk-test.jsp

Table 2: Variable List Diagreement for CVD



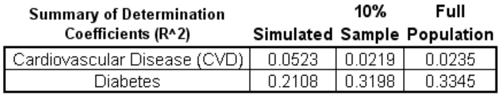
Variable	HSPH	NIH/AHA
Are you diabetic?	X	
Diastolic blood pressure	X	
Do you have a sibling or parent that had a heart attack?	X	
Do you smoke? (Yes, No, Quit)	X	X
(Yes) Number of cigarettes smoked per day	X	
(Quit) How long ago did you quit smoking?	X	
Do you take a multivitamin most days?	X	
Do you take vitamin B-complex suppliments most days?	X	
Do you take vitamin E suppliments most days?	X	
Do you walk 30 minutes per day?	X	
Gender	X	X
HDL cholesterol level	X	X
Height in inches	X	
How often are you exposed to second-hand smoke?	X	
Servings of alcohol per day	X	
Servings of fish per week	X	
Servings of fruits per day	X	
Servings of grains per day	X	
Servings of nuts per week	X	
Servings of saturated fats per day	X	
Servings of unsaturated fats per day	X	
Servings of vegetables per day	X	
Systolic blood pressure	X	X
Total cholesterol level	X	X
Waist in Inches	X	
Weight in pounds	X	
Age		X
Are you taking medications for high blood pressure?		X

Table 3: Variable List Diagreement for Diabetes



Variable	HSPH	ADA
Do you have a parent with diabetes?	X	X
Do you have a sibling with diabetes?	X	X
Do you smoke? (Yes, No, Quit)	X	
(Yes) Number of cigarettes smoked per day	X	
(Quit) How long ago did you quit smoking?	X	
Do you walk 30 minutes per day?	X	X
Gender (Male, Female)	X	
(Female) Have you had gestational diabetes?		X
Height in inches	X	X
Servings of alcohol per day	X	
Servings of grains per day	X	
Servings of starch per day	X	
Servings of unsaturated fats per day	X	
Waist size in inches	X	
Weight in pounds	X	X
What is your racial association?	X	
Age		X

Table 4



Summary of Determination Coefficients (R^2)	Simulated	10% Sample	Full Population
Cardiovascular Disease (CVD)	0.0523	0.0219	0.0235
Diabetes	0.2108	0.3198	0.3345

4. Modeling Platform and Approach

As has been described previously in [4, 5, 6, 7], the PROCEED™ process is a recipe-like approach applicable to a wide variety of artificial intelligence and data mining tasks. While originally developed to solve challenging problems in manufacturing optimization, several concepts hold true between the manufacturing and medical applications.

[2] Generating valid simulations of real human populations is very challenging and will be addressed in a later paper.

Figure 1: Frequency Scatterplot - Public CVD Model Comparison

Simulated Patient Data, 3479 Undiagnosed Cases

Uniform Relative Scaling Applied (0% - 100% full-scale)

AHA/NIH CVD - Rescaled : HSPH CVD - Rescaled: $r^2 = 0.0523$

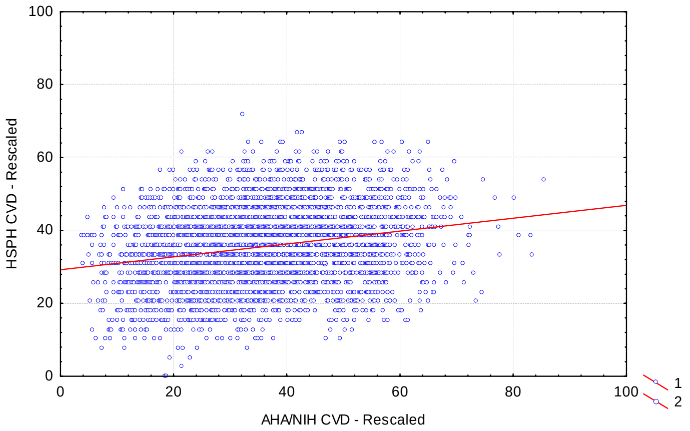


Figure 2: Frequency Scatterplot - Public Diabetes Model Comparison

Simulated Patient Data, 3479 Undiagnosed Cases

Uniform Relative Scaling Applied (0% - 100% full-scale)

ADA Diabetes - Rescaled : HSPH Diabetes - Rescaled: $r^2 = 0.2108$

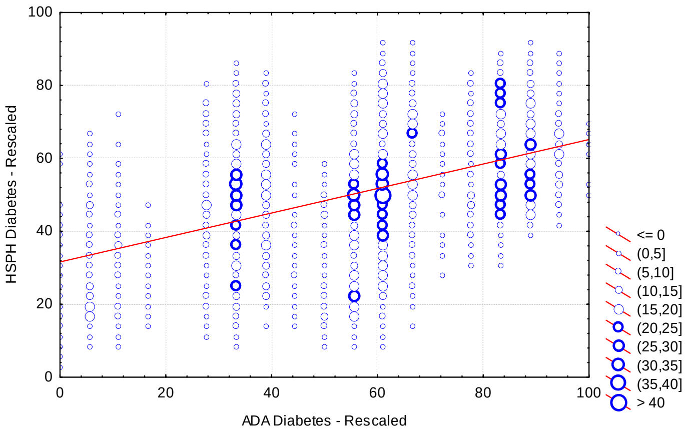


Figure 3: Frequency Scatterplot - Public CVD Model Comparison

2005 Patient Data, 8679 Undiagnosed Cases

Uniform Relative Scaling Applied (0% - 100% full-scale)

2005 AHA/NIH CVD - Rescaled : 2005 HSPH CVD - Rescaled: $r^2 = 0.0219$

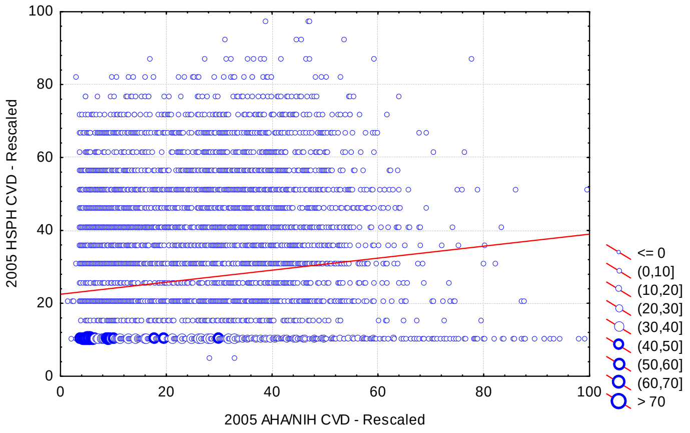


Figure 4: Frequency Scatterplot - Public Diabetes Model Comparison

2005 Patient Data, 12842 Undiagnosed Cases

Uniform Relative Scaling Applied (0% - 100% full-scale)

2005 ADA Diabetes - Rescaled : 2005 HSPH Diabetes - Rescaled: $r^2 = 0.3198$

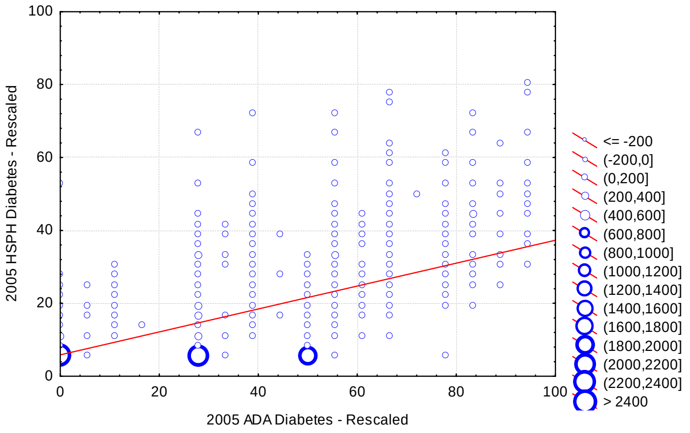


Figure 5: Frequency Scatterplot - Public CVD Model Comparison

2005 Patient Data, 91850 Undiagnosed Cases

Uniform Relative Scaling Applied (0% - 100% full-scale)

2005 AHA/NIH CVD - Rescaled : 2005 HSPH CVD - Rescaled: $r^2 = 0.0235$

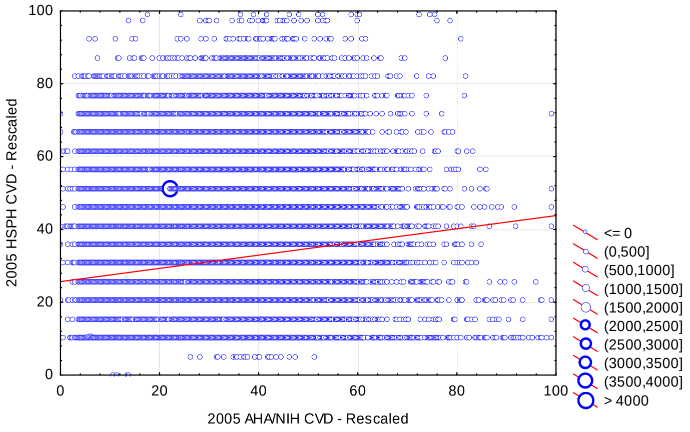
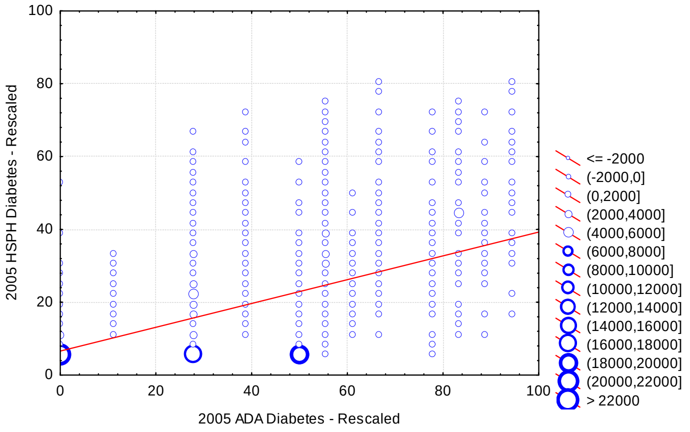


Figure 6: Frequency Scatterplot - Public Diabetes Model Comparison

2005 Patient Data, 128876 Undiagnosed Cases

Uniform Relative Scaling Applied (0% - 100% full-scale)

2005 ADA Diabetes - Rescaled : 2005 HSPH Diabetes - Rescaled:: $r^2 = 0.3345$



"First do no harm" is a key element of our operating philosophy. The PROCEED modeling process is biased is such a way that it will either produce a high-quality model or deliver no model at all. Deploying an incorrect model engenders unacceptable risks in both environments. PROCEED embodies an expert system approach to artificial intelligence, guiding the user away from deployment of a potentially unqualified model.

Both manufacturing and medicine produce massive quantities of data. Organizing this data into information and subsequently extracting knowledge from it are extremely challenging due to the data volumes involved. Discovering which variables matter most in a particular outcome requires careful detection of very high order interactions, most of which can not be discovered through conventional approaches like designed experiments. The Mahalanobis Distance Genetic Algorithm (MDGA) process [6] is particularly valuable in rapidly identifying potential variable combinations that relate to one another in complex ways.

Finally, once a well-qualified model is discovered and validated we need to understand what actions to take to provide the most beneficial outcome. Too often a naive approach is applied where recommendations take the form of an "all or nothing" proposition – change completely and without variation or don't bother changing at all. What we desire instead is a deeper understanding that identifies what must be controlled and at the same time considers what controls may be relaxed with little or no penalty. From the standpoint of applying MRS models in public health environments, this translates to knowing where we should focus efforts and funding to drive population change and at the same time being aware of opportunities to reduce emphasis that will have little net benefit.

5. Resolving Conflicts

The process for resolving conflicts can be most easily described in two dimensions and is illustrated in Figure 7. We know that the pairs of models agree along a line of slope 1. That is, at certain points both models agree on a risk score when considered in FSR% terms. Moving off this line, we can imagine circular arcs of equal risk passing though the strata. At the extreme, one model may indicate 100% FSR% and the other 0%, resulting in a radial FSR% of 100%. In the extreme when both models indicate 100% FSR%, we calculate a radial FSR% of 141.42%. As needed, we can normalize this new radial space to a 0% - 100% FSR% scale as before. Likewise other combinations along the continuum of pairs can be represented[3].

Once the multiple models are converted into a single resolved score, we apply the PROCEED process to its prediction. We include the superset of variables as potential inputs, then allow the PROCEED process to reduce the variables, model the relationships precisely and validate them through stochastic simulation. The architecture of the resulting hybrid models for CVD and diabetes are shown in Table 5. Both are conventional multi-layer perceptron (MLP) models [8] with a single hidden layer. Activation functions differ by layer and are also reflected in the table.

[3] This approach can and has been scaled up to more than two dimensions, resulting in radii of hyperspheres aligning more than three models. If the maximum FSR% of the contributing models is 100%, then the maximum expected radial FSR% is $(\textit{number of models} * \textit{FSR\%}_{max}{}^{2})^{0.5}$. Future papers will document these results.

All neurons in a particular layer have the same activation function

6. Resolution quality

How would one know if one MRS model is “better” than another? The most direct approach is to observe which model best predicts disease onset. Unfortunately most public MRS models produce relative scores, not absolute measures of risk. We can resolve this by applying Bayesian Posterior Probability (BPP) [9] to the risk stratification results.

BPP is defined as:

$$p(A_t \mid X_{\psi,t}) \equiv \left[\frac{p(X_{\psi,t} \mid A_t) \cdot p(A_t)}{p(X_{\psi,t} \mid A_t) \cdot p(A_t) + p(X_{\psi,t} \mid \sim A_t) \cdot p(\sim A_t)} \right]$$

when,

$$\psi \equiv \bar{x}_{\hat{A}_t} + n\sigma_{\hat{A}_t}$$

where A_t is a positive diagnosis of the disease at the moment in time t and $X_{\psi,t}$ is the score of that individual being greater than ψ at the moment in time t.

The ψ calculation is repeated to determine an alarm threshold specific to each MRS model, thereby allowing each model to determine it’s own cutoff for

statistically significant high-risk populations. We define $n=2$ for our purposes. We then obtained the risk scores for the real-world population of the Healthy Balance program in 2004, and identified patients that became positive for either CVD or diabetes in 2005. This satisfies the Bayesian requirement for three identifiable populations:

$X_{\psi,t} \mid A_t$ - those positive given their score is above the threshold $X_{\psi,t}$

$X_{\psi,t} \mid \sim A_t$ - those negative or undiagnosed given their score is above the threshold $X_{\psi,t}$

A_t - those positive in the general population

Figure 7: Radial Hybridizaion Example
Simulated Patient Data, 3479 Undiagnosed Cases
Uniform Relative Scaling Applied (0% - 100% full-scale)

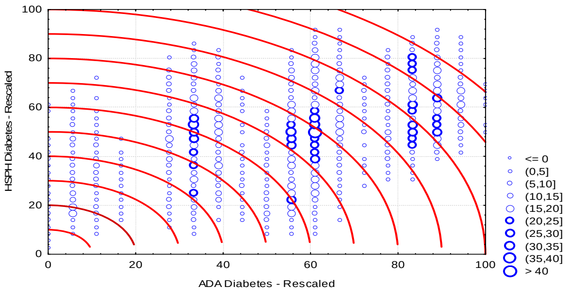
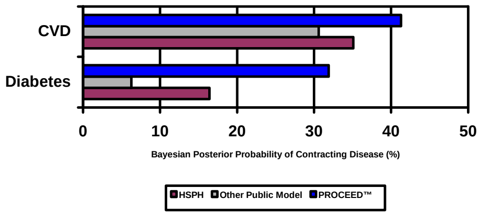


Table 5: Architctures of Hybrid MLP Models

	Input Layer		Hidden Layer		Output Layer	
	# Neurons	Activation Function	# Neurons	Activation Function	# Neurons	Activation Function
Cardiovascular Disease (CVD)CVD	73	Sigmoid	41	Tanh	1	Tanh
Diabetes	76	Sigmoid	14	Tanh	1	Identity

Figure 9 displays the BPP results for the pairs of public risk models and compares them to the resulting PROCEED hybrid model. In both the CVD and diabetes cases the PROCEED-based hybrid outperforms the source public MRS models by a significant margin when tested on the complete Healthy Balance population.

Figure 9: Absolute Disease Risk Comparison
2005 Patient Data, Measuring Detection Capability 1 Year Prior to Disease Onset
Bayesian Posterior Probaility (BPP) Based on +2 Standard Deviation Alarm Level



7. Conclusion & Future Work

MRS models are highly useful public health management tools. However, disagreement in public MRS models is problematic for those tasked with investing limited funds and resources to support a large population. By converting such models to a common FSR% basis then fitting models to radial strata of their intersections, we have increased the BPP of our identification and improved our ability to address high-risk persons who may contract CVD or diabetes in the future while they still have to opportunity to avoid such an outcome.

More complex diseases, such as cancer, can have many manifestations and pathologies. One part of our ongoing work includes combining models of similar but not identical pathologies to create effective yet more general models of these challenging conditions.

As was discussed previously, we desire an intervention selection process that balances factors that require tighter control and increased emphasis with identification of those factors that can be relaxed with little or no penalty. We must also consider the performance of MRS models and related recommendations over time. In particular, we desire models that provide as much accuracy and precision as possible as early as possible before disease onset. This would allow public health programs more time to act on their populations to further improve health. Solving the problems of making predictions and measuring MRS model predictive accuracy and precision over time will be discussed in a later paper.

Finally, as data warehouses grow we gain the ability to calculate trajectories of risk over time. To improve the effectiveness of our interventions we would like to identify both those at high risk now and also those whose trajectory indicates a rapid rise in risk that will eventually lead to an unpleasant outcome. Work continues on these trajectory-based techniques.

8. References

[1] “Postmenopausal Hormone Therapy and Cardiovascular Disease in Women,” http://www.americanheart.org/presenter.jhtml?identifier=4536, American Heart Association, as viewed February 2007.

[2] “Research Milestones.” http://www.nhlbi.nih.gov/about/framingham/timeline.htm, National Heart, Lung and Blood Institute, as viewed February 2007.

[3] Healthy Balance Data Privacy Statement, http://cathealthbenefits.cat.com/cda/components/securedFile/displaySecuredFileServletJSP?fileId=189489&languageId=7, as viewed February 2007.

[4] A. Grichnik and M. Seskin, “Process Modeling: Challenges and Applications”, in Proceedings of the 5th International Conference on Recent Advances in Soft Computing (RASC 2004), Nottingham, England, 2004.

[5] B. Clarke, K. Dove, A. Grichnik, M. Seskin, D. Weis, U.S Patent Publication US20050055176A1 - Method of analyzing a product. (See also WO06085877A1.) 2005.

[6] A. Grichnik, M. Seskin., U.S Patent Publication US20060230018A1 - Mahalanobis distance genetic algorithm (MDGA) method and system. (See also WO06110244A2.) 2006.

[7] V. Bhasin, A. Grichnik, M. Seskin, US Patent Publication US20060229852A1, Zeta statistic process method and system. (See also WO06110242A2.) 2006.

[8] F. Rosenblatt, “The perceptron: a probabilistic model for information storage in the brain”, *Psychological Review* 65, pp 385-408. American Psychological Association, 1958.

[9] E.Yudkowsky, “An Intuitive Explanation of Bayesian Reasoning”, http://yudkowsky.net/bayes/bayes.html, as viewed February 2007.